METHODES STATISTIQUES POUR L'ETUDE DE LA STRUCTURATION SPATIALE DE LA DIVERSITE GENETIQUE

Pierre Faubet

UNIVERSITE GRENOBLE I - JOSEPH FOURIER
ECOLE DOCTORALE INGENIERIE POUR LA SANTE, LA COGNITION ET L'ENVIRONNEMENT

Modèles, méthodes et algorithmes en biologie, santé et environnement

Pierre FAUBET

METHODES STATISTIQUES POUR L'ETUDE DE LA STRUCTURATION SPATIALE DE LA DIVERSITE GENETIQUE

sous la direction d'Oscar E. GAGGIOTTI

Manuscrit soumis au jury composé de

Frédéric AUSTERLITZ (rapporteur), Chargé de recherche CNRS, Paris
Frantz DEPAULIS (examinateur), Chargé de recherche CNRS, Paris
Olivier FRANCOIS (examinateur), Professeur, Institut National Polytechnique de Grenoble
Oscar GAGGIOTTI (directeur de thèse), Professeur, Université Grenoble I
Xavier VEKEMANS (rapporteur), Professeur, Université de Lille I

Thèse préparée au Laboratoire d'ECologie Alpine, UMR CNRS 5553

Titre : Méthodes statistiques pour l'étude de la structuration spatiale de la diversité génétique.

Résumé
La sélection naturelle et les flux de gènes entre populations contribuent à structurer la diversité génétique dans l'espace sous l'influence de l'environnement. L'étude de ces forces évolutives et de leur interaction avec le milieu a des applications importantes dans des domaines tels que la biologie de la conservation, la génétique ou l'agronomie. Les données génétiques peuvent être reliées aux données environnementales à travers des modèles qui décrivent les processus évolutifs mis en jeu pour estimer des paramètres d'intérêt. Le développement d'une méthode d'estimation en génétique des populations consiste donc à construire un modèle selon des considerations biologiques pour l'utiliser ensuite dans des algorithmes d'estimation. L'étape suivante consiste alors à évaluer les performances de la méthode pour la valider ou l'améliorer. Ce schéma a été appliqué pour évaluer une méthode d'estimation des taux de migration qui a été étendue par la suite. Une autre méthode a été développée pour étudier l'adaptation locale sous l'influence de la migration et de la sélection naturelle.

Mots-clés : génétique des populations, migration, sélection naturelle , modèles bayésiens hiérachiques, méthodes d'inférence statistique.

Title: Statistical methods for the study of the spatial structure of genetic diversity.

Abstract
Natural selection and gene flows between populations shape the spatial distribution of genetic diversity and are influenced by environmental factors. The study of these evolutionary forces and their environmental origins finds important applications in the fields of conservation biology, genetics or agriculture. Genetic and environmental data can be combined in models that describe considered evolutionary processes to estimate parameters of interest. Developping statistical genetic methods consists of establishing a model based on biological considerations and, then, using this model to get estimations. The next step consists of evaluating such methods to validate its performances or to improve it. This scheme was applied to evaluate a method for the estimation of migration rates that was subsquently extended. Another method was developped for the study of local adaptation due to migration and natural selection.

Keywords: population genetics, migration, natural selection, hierarchical bayesian models, statistical inference methods.

Table des matières

La structure spatiale de la diversité génétique est influencée par la migration entre populations et la sélection naturelle. Son étude a des applications importantes dans le domaine de la conservation, de l'amélioration des espèces agricoles, et de l'identification des gènes impliqués dans le déterminisme de maladies génétiques ou dans la résistance à des pathogènes. Une bonne compréhension de la structure génétique des populations est fondamentale dans ces applications. La compréhension des facteurs qui influencent la structure génétique des populations peut également aider à concevoir des stratégies de gestion adaptées pour la conservation de la diversité génétique. L'objectif de la thèse est de développer des méthodes statistiques qui permettront d'évaluer le rôle de la migration et de la sélection naturelle dans la répartition spatiale et la diversité génétique. Les paramètres d'intérêt seront estimer selon une approche bayésienne hiérarchique. Ce travail, de nature pluridisciplinaire, nécessite des compétences en biologie, en mathématiques et en informatique. Les concepts de base en génétique des populations et en statistique bayésienne seront présentés afin de rendre ce travail accessible au plus grand nombre.

Ce travail comprend trois parties :

- La première, introductive, présente et illustre la notion de structure spatiale de la diversité génétique. Une description de l'approche bayésienne hiérarchique et de son intérêt en génétique des populations vient compléter cette partie.
- La deuxième partie est consacrée à l'estimation des taux de migration et à l'influence des facteurs environnementaux sur les processus de dispersion. Les performances d'une première méthode sont évaluées à partir de données simulées. Au vu des résultats une nouvelle méthode est développée.
- La troisième partie traite des variations de fréquences alléliques le long de gradients environnementaux. Une méthode d'estimation des coefficients de sélection est proposée.

Enfin une discussion générale présente les conclusions de cette étude.

Première partie

Introduction

Chapitre 1

La structuration spatiale de la diversité génétique

Ce chapitre introduit les mécanismes responsables de la distribution géographique de la diversité génétique. Diverses applications illustrent l'importance de l'étude de la structure spatiale des gènes et des facteurs environnementaux qui l'influencent.

1.1 Quelques notions et illustrations

Avant d'expliquer l'intérêt de l'étude de la structuration spatiale de la diversité génétique, il convient d'en définir les objets : la diversité génétique, sa structure, sa distribution dans l'espace.

1.1.1 La diversité génétique : un niveau de biodiversité

L'ensemble des caractères apparents des individus d'une espèce, leur phénotype, résulte de l'interaction de leur patrimoine génétique, leur génotype, avec leur environnement. La pigmentation de la peau chez l'homme est parmi les exemples les plus parlant de tels caractères. Ainsi, au sein d'une espèce, derrière la richesse des phénotypes se cache un niveau de variété des gènes appelée *diversité génétique* ou *diversité intraspécifique*.

La diversité génétique se caractérise par :

- les différentes formes d'un gène (les *allèles*) observables à un endroit donné du génome (un *locus*),
- la fréquence des allèles en question,
- la fréquence des génotypes au locus considéré.

La richesse en allèles révèle le *polymorphisme* qui est à l'origine de la diversité des génotypes et des phénotypes. Les fréquences alléliques et génotypiques permettent de décrire la composition génétique d'une population.

L'observation des groupes sanguins chez l'homme illustre ces différents niveaux de diversité. Le locus concerné présente trois allèles (A, B et O) qui peuvent former six génotypes (A/A, A/O, B/B, B/O, A/B et O/O) et quatre groupes sanguins (phénotypes A, B, AB et O). Les proportions de chaque allèle et de chaque génotype déterminent la fréquence des différents groupes sanguins.

1.1.2 La structure génétique des populations

L'ensemble des fréquences alléliques et génotypiques constitue la *structure génétique des populations* qui permet de décrire la diversité intraspécifique au locus considéré. La comparaison des structures génétiques en différents loci sert à mesurer le niveau d'association des allèles entre différentes régions du génome (déséquilibres gamétiques).

Cette étude se limite au seul cas des espèces diploïdes, celles pour lesquelles les allèles sont associés par paire dans chaque individu (un issu de chaque parent). La connaissance de la structure génotypique donne directement la structure allélique mais le passage inverse est impossible sans informations ou hypothèses supplémentaires. En effet, les mêmes fréquences alléliques peuvent produire des structures génotypiques très différentes selon le mode d'association des allèles lors de la reproduction : par exemple, une population avec 50% de gènes A et 50% de gènes a peut ne contenir que des homozygotes AA et aa, ou que des hétérozygotes Aa.

La génétique des populations s'intéresse à l'évolution de la fréquence des gènes dans les populations et se focalise donc sur la structure allélique. La population est vue comme un ensemble de gènes portés par les individus et dont la composition est susceptible de changer au fil des générations sous l'influence de pressions évolutives (cf. §1.2). Cette évolution temporelle de la structure génétique dépend fortement de l'environnement dans lequel ce processus se déroule.

1.1.3 La structuration spatiale de la diversité génétique

Les variations spatiales de la structure génétique des populations à travers un territoire résultent du mode d'occupation de l'espace et/ou de chagements des conditions environnementales locales. En effet, selon les espèces, les individus sont répartis de façon continue dans leur habitat ou s'organisent en groupes ou sous-populations. Par ailleurs, dans un environnement hétérogène, certains individus sont mieux adaptés que d'autres pour survivre dans un milieu donné.

L'organisation en populations discrètes peut produire des structures génétiques contrastées selon le degré d'isolement des sous-populations. Ce mode d'occupation

de l'espace s'observe chez les espèces sociales ou provient de la fragmentation de l'habitat. Moins il y a de connexions entre les populations, plus il y a de chances que leurs histoires évolutives divergent. Le phénomène peut s'amplifier si, de plus, les conditions environnementales locales changent dans le temps et dans l'espace.

Diverses études ont mis en évidence la structure spatiale de la diversité génétique à différentes échelles, chez un grand nombre d'espèces. Par exemple, chez l'homme, les travaux de Novembre et al. (2005) sur une mutation qui confère une résistance au VIH s'intéressent aux changements de fréquence de cette mutation à travers l'Europe et l'Asie (cf. Figure 1.1(a)). Les auteurs de l'étude ont pu mesurer l'intensité de la sélection et estimer l'origine géographique de la mutation sous divers scénarios environnementaux. D'autres recherches, menées par Rosenberg et al. (2002), portent sur la structure génétique des populations humaines (cf. Figure 1.1(b)). L'analyse d'un échantillon de gènes de 1056 individus issus de 52 populations sur le globe, leur a permis d'identifier six groupes génétiques. Cinq d'entre eux correspondent aux principales régions géographiques (Afrique, Eurasie, Asie de l'est, Océanie et Amériques).

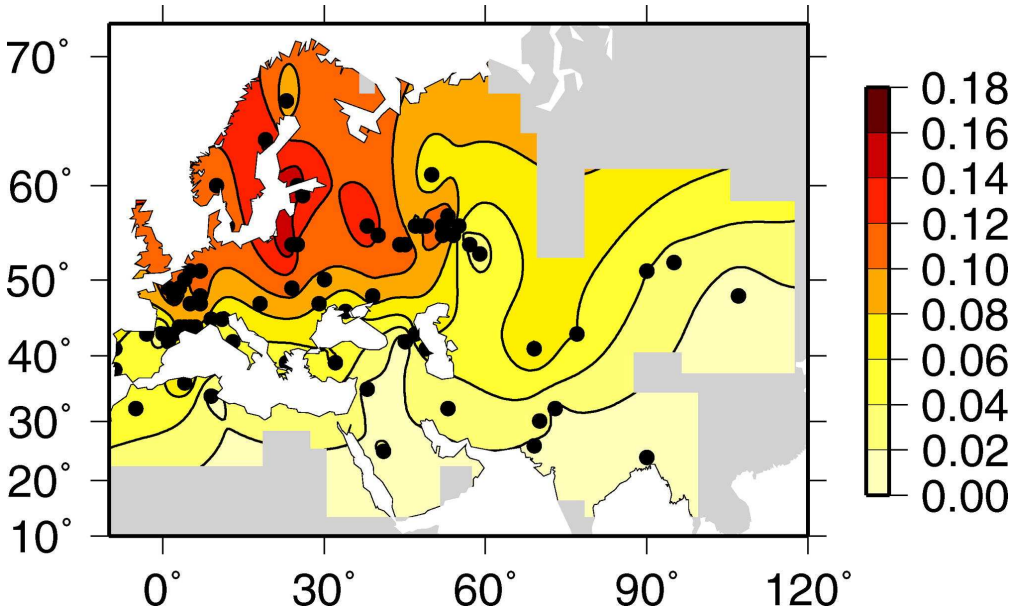
1.2 Mécanismes évolutifs et environnement

Pour mieux comprendre comment l'environnement influence la structure spatiale de la diversité génétique, il convient d'introduire les mécanismes qui en gouvernent l'évolution. Les effets conjoints des processus évolutifs décrits ci-après expliquent la diversité génétique observée dans les populations naturelles.

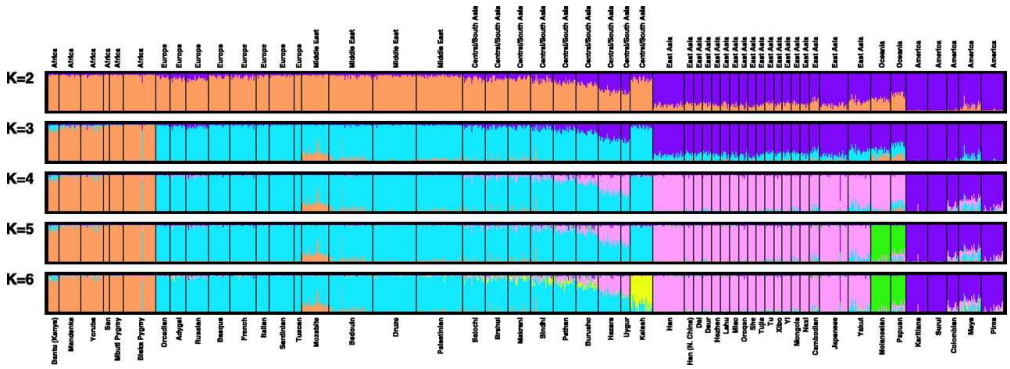
Il existe quatre types de pressions, ou forces, évolutives qui agissent sur les fréquences alléliques d'une population ; la *dérive génétique*, les *mutations*, les *flux de gène*s et la *sélection naturelle* modifient la structure génétique au fil des générations. Leurs effets sur la diversité génétique à un locus sont présentés ci-après. Le cas multilocus sera abordé par la suite.

La dérive génétique

Dans une population d'effectif fini les gènes d'une génération constituent un échantillon des gènes de la génération précédente. L'effet de cet échantillonnage lors de la reproduction, la dérive, induit des variations aléatoires de la fréquence des gènes au fil des générations. Cumulée sur de nombreuses générations cette force aboutit à la fixation de l'un des deux allèles (et à la perte de l'autre) et diminue ainsi le niveau de diversité génétique. La dérive génétique est liée à la taille de la population, elle sera d'autant plus forte que l'effectif est réduit. Ainsi, la dérive génétique n'est pas (directement) influencée par les conditions environnementales, mais par l'histoire démographique de la population étudiée.



(a) Novembre et al. 2005, The geographic spread of the CCR5 Delta32 HIV-resistance allele. Cartes des variations de la fréquence d'un allèle.



(b) Rosenberg et al. 2002, Genetic structure of human populations. Constitution de groupes d'individus génétiquement proches.

FIGURE 1.1 – Deux exemples de la structuration spatiale de la diversité génétique chez l'homme.

Les mutations

Des modifications accidentelles du matériel génétique, transmises d'une génération à la suivante, produisent de nouveaux allèles. Le rôle des mutations dans l'évolution est primordial, à l'origine de la variabilité génétique elle contribue à augmenter la diversité génétique. Son influence sur la fréquence des gènes est négligeable par rapport à celle des autres forces qui détermineront le devenir des mutations. Bien que l'environnement influence parfois ces évènements (rayons UV, radioactivité, ...), de telles modifications sont accidentelles et limitées dans le temps.

Les flux de gènes

Les processus de dispersion des individus d'une espèce (colonisation et migration) conduisent à des échanges de matériel génétique entre individus issus de différentes populations. Au cours du temps, cette force tend à homogénéiser les fréquences alléliques des populations et contribue à maintenir la diversité génétique.

Les flux de gènes dépendent de la mobilité des individus au sein de leur habitat et par conséquent des facteurs environnementaux. A titre d'exemples, les paysages accidentés, a priori, limitent les déplacements et les échanges génétiques alors que la proximité géographique les favorisent. Les conditions du milieu influencent donc la migration et la structure spatiale des fréquences alléliques.

La sélection naturelle

Le processus à travers lequel les individus les mieux adaptés à leur environnement survivent et se reproduisent conduit à des modifications de la structure génétique. Ainsi les allèles qui confèrent un avantage sélectif voit leur fréquence augmenter au fil des générations.

En fonction des conditions environnementales, un allèle peut très bien être favorisé à un endroit donné et défavorisé dans un autre. Ainsi les changements de conditions environnementales contribuent à faire apparaitre des variations spatiales des pressions sélectives et de la structure des fréquences alléliques.

Génétique à plusieurs loci, recombinaison et déséquilibre de liaison

Pour mieux appréhender la diversité génétique il peut être nécessaire de considérer plusieurs loci simultanément. Dès lors, la connaissance de la structure génétique à chaque locus ne suffit pas, il faut aussi tenir compte de l'association des allèles à différents loci.

Les structures alléliques de chaque locus évoluent non seulement sous l'influence des pressions évolutives décrites ci-dessus mais aussi sous l'effet de la *recombinaison*. En effet, les échanges de matériel génétique entre deux fragments d'ADN assure le brassage génétique par la formation de nouvelles combinaisons génétiques. Ainsi, la recombinaison tend à diminuner les associations non aléatoires entre allèles de loci différents, les *déséquilibres de liaison* ; elle contribue ainsi à augmenter la variabilité génotypique et à maintenir la diversité génétique.

Les causes d'apparition des déséquilibres de liaison peuvent être purement démographique (e.g. mélange de populations) ou liées à des pression évolutives (e.g. auto-stop génétique). Dans les deux cas la recombinaison tend à rétablir l'équilibre.

Il apparaît que les modifications de la structure spatiale de la diversité génétique dues à la migration et à la sélection naturelle sont liées aux facteurs environnementaux. L'identification de la structure des populations naturelles passent donc par l'étude de ces processus et des variables environnementales qui les influencent.

1.3 Applications et motivations

L'étude de la structure génétique trouve des applications importantes dans la gestion et l'amélioration des espèces, dans l'identification de gènes sous sélection. Ces problèmes d'interactions gènes/environnement requièrent la connaissance de la distribution spatiale des gènes.

Dans le domaine de la biologie de la conservation, connaître la structure génétique permet d'identifier des unités de gestion qui seront la cible des mesures de conservation. Par ailleurs, la compréhesion des facteurs qui influencent la distribution spatiale des gènes sert à concevoir des stratégies de gestion adaptées selon l'environnement. Les diagnostics issus de ces applications permettent aux gestionnaires de faire des choix éclairés sur les actions à engager pour le maintien de la biodiversité.

Dans le domaine de l'agronomie, l'identification de gènes de résistance ou de sensibilité à certaines maladies requiert l'analyse de la structure génétique. La détection de gènes responsables de l'adaptation à certaines contraintes environnementales (pollution, sécheresse, ravageurs, ...) nécessite le même type d'analyses. Ainsi l'amélioration des récoltes et des espèces agricoles passe par une meilleure compréhension de la distribution de la diversité génétique.

Chez l'homme, l'étude de la structure génétique permet l'identification de gènes impliqués dans le déterminisme de maladies génétiques (mucoviscidose, anémie, ...). Il en est de même en épidémiologie pour les gènes de résistance à des pathogènes (VIH, malaria, ...). Dans ces cas la connaissance de la structure génétique fournit des informations sur les conditions environnementales qui ont influencé

l'évolution de ces gènes.

Plus généralement l'étude de la structuration spatiale de la diversité génétique permet de mieux comprendre

- l'adaptation des espèces à leur environnement,
- les processus d'apparition de nouvelles espèces, la *spéciation*, en fonction des contraintes du milieu.

Ainsi, l'étude de l'influence des facteurs environnementaux sur la distribution spatiale de la diversité génétique est un préalable indispensable à toutes les applications mentionnées plus haut.

1.4 Objectifs de la thèse

Le but de cette étude est de développer des méthodes statistiques qui permettront :

- d'estimer les taux de migration entre population et de mesurer l'influence de l'environnement sur la migration,
- de déterminer les facteurs responsables de l'adaptation locale.

Les méthodes développées combineront données génétiques et environnementales pour estimer les paramètres de modèles bayésiens hiérarchiques par des algorithmes numériques intensifs. Cette approche s'est révélée particulièrement adaptée aux études de génétique des populations qui impliquent la modélisation de processus évolutifs complexes.

Chapitre 2

Inférence bayésienne en génétique des populations

Ce chapitre présente les bases de l'inférence bayésienne et les raisons de son utilisation pour traiter différentes questions en génétique des populations.

Durant son développement la recherche en génétique des populations a tenté d'expliquer l'évolution des populations naturelles par des modèles théoriques. Ce travail fut initié par les pères de la discipline, R.A. Fisher, J.B.S. Haldane et S. Wright, dans les années 1920-1930. Longtemps freinée par le manque de données la génétique des populations profite aujourd'hui de nouveaux outils. D'une part les récentes avancées des techniques moléculaires mettent à disposition des chercheurs une quantité considérable de données. D'autre part, les moyens de calcul qui permettent le traitement des données et la simulation des modèles n'ont cessé de progresser.

La génétique des populations repose sur l'élaboration de modèles mathématiques qui, confrontés à la réalité, permettent de mesurer l'influence des pressions évolutives. Les forces qui agissent sur l'évolution des fréquences alléliques sont intrinsèquement stochastiques et s'étudient par des modèles probabilistes. Les recherches en génétique des populations ont d'abord porté sur l'évolution des fréquences alléliques. Par la suite, deux types d'approches ont permis de construire les modèles : le coalescent (Kingman 1982a, 1982b) à partir de la reconstruction de la généalogie des gènes, l'approche génotypes multilocus à partir des déséquilibres gamétiques. La complexité des processus et des scénarios considérés, la quantité de paramètres interdépendants des modèles requièrent une approche adaptée à ce type de problèmes.

L'inférence bayésienne, par ses aspects pratiques et sa flexibilité, permet de traiter diverses questions en génétique des populations. L'approche bayésienne a permis de développer des méthodes pour, entre autres, étudier la structure génétique des populations (e.g. TESS, François et al. 2006 ; STRUCTURE, Pritchard et

al. 2000), évaluer l'influence des facteurs environnemetaux sur la colonisation (e.g. COLONIZE, Gaggiotti et al. 2004), sur la différentiation génétique (e.g. GESTE, Foll and Gaggiotti 2006), estimer les taux de migration entre populations (e.g. BAYESASS, Wilson and Rannala 2003 ; MIGRATE, Beerli 2006). D'autres applications concernent l'estimation de paramètres démographiques chez *Arabidopsis thaliana* (François et al. 2008), chez l'homme (Fagundes et al. 2007), chez le bacille de Koch (Tanaka et al. 2006), chez le campagnol (Hamilton et al. 2005), chez l'abeille (Excoffier et al. 2005).

La suite du chapitre propose une introduction à l'inférence bayésienne. La finalité de l'inférence est de prévoir et de fournir des outils d'aide à la décision aux gestionnaires et aux chercheurs. Ainsi, l'objectif est de déterminer des estimateurs ponctuels (moyenne, mode, médiane, ...) et/ou ensemblistes (intervalles de confiance) des paramètres d'un modèle. L'intérêt de la statistique bayésienne par rapport à la statistique classique est qu'elle permet également d'estimer la distribution a posteriori des paramètres. Le modèle peut lui aussi être l'objet de l'inférence lorsqu'il s'agit de tester sa validité ou de le comparer à d'autres. L'estimation de la fréquence p d'un allèle A à un locus biallélique permettra d'illustrer la démarche.

2.1 Statistique bayésienne

Le but de l'*inférence bayésienne* est d'estimer le(s) paramètre(s) θ d'un modèle probabiliste à partir de données $\mathcal{D}$ issues de l'observation. Le modèle constitue une représentation mathématique du phénomène étudié, les données sont considérées comme la réalisation de variables aléatoires dont la distribution dépend du(es) paramètre(s) du modèle.

En statistique bayésienne, par opposition à l'approche classique (ou *fréquentiste*), les paramètres sont vus comme des variables aléatoires. Cette considération introduit une part de subjectivité via la distribution a priori des paramètres, $\Pr(\theta)$, qui ne tient pas compte des données. L'inférence bayésienne consiste à confronter cet a priori aux observations pour calculer la *distribution a posteriori* des paramètres sachant les données, $\Pr(\theta|\mathcal{D})$.

Bien que controversée, l'utilisation d'un a priori permet d'apporter l'information de modèles élaborés lorsque peu de données sont disponibles. Par ailleurs, le choix de l'a priori restreint ou contraint l'espace des paramètres pour ne s'intéresser qu'aux valeurs qui semblent les plus réalistes.

L'estimation de la fréquence p de l'allèle A dans un échantillon de $N = 10$ individus (diploïdes) permet d'illustrer la démarche. Dans cet exemple le paramètre θ à estimer est la fréquence p, la donnée $\mathcal{D}$ est $k = 6$, le nombre d'occurrences de l'allèle sur $2N = 20$ allèles observés. L'estimation naturelle (et fréquentiste) de

la fréquence de A est le rapport du nombre d'occurrences de A au nombre total d'observations, soit $k/2N = 6/20 = 0,3$.

La première étape de la démarche consiste à établir la *vraisemblance* des données sachant la valeur du paramètre du modèle probabiliste, $\Pr(\mathcal{D}|\theta)$. La fonction $\Pr(\mathcal{D}|\theta)$ rend compte du lien stochastique entre les données et les paramètres du modèle. Connaissant la fréquence p de l'allèle A, la probabilité d'en observer k copies parmis $2N$ suit une loi binomiale $\mathcal{B}(2N, p)$:

$$\Pr(\mathcal{D}|\theta) = \binom{2N}{k} p^k (1-p)^{2N-k}$$

Dans ce cas, les données sont supposées indépendantes et identiquement distribuées (i.e. de même loi). Intuitivement, le produit $p^k(1-p)^{2N-k}$ est la probabilité d'obtenir une combinaison de k copies de A parmis $2N$ allèles et le coefficient $\binom{2N}{k}$ est le nombre de combinaisons possibles de k allèles A parmis $2N$.

Le choix de la distribution a priori $\Pr(\theta)$ peut se faire selon des considérations théoriques (modélisation) ou pratiques (i.e. pour faciliter les calculs). Quoi qu'il en soit la loi a priori ne tient pas compte des données, elle révèle le niveau de connaissance du phénomène étudié. Pour l'estimation de p, il est possible que la population actuelle soit issue d'une population ancestrale dans laquelle la fréquence de l'allèle était $1/2$. Le niveau de différentiation génétique entre les deux populations (distance génétique) est supposé égal à 20. Sous ces hypothèses, les travaux de Wright (1931) montrent que la fréquence de l'allèle suit une loi de type Beta centrée en $1/2$ (cf. courbe bleue Figure 2.1(a)),

$$p \backsim \beta(\alpha, \alpha), \quad \text{i.e.} \quad \Pr(\theta) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} p^{\alpha-1}(1-p)^{\alpha-1}, \quad p \in (0,1)$$

avec $\alpha = 20 \times 1/2 = 10$.

La deuxième étape consiste à établir l'expression de la distribution a posteriori $\Pr(\theta|\mathcal{D})$ à partir de la formule de Bayes

$$\Pr(\theta|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\theta)\Pr(\theta)}{\Pr(\mathcal{D})} \tag{2.1}$$

avec

$$\Pr(\mathcal{D}) = \int \Pr(\mathcal{D}|\theta)\Pr(\theta)\, d\theta \tag{2.2}$$

L'application à l'estimation de la fréquence de l'allèle A démontre que la loi a posteriori de p est de type Beta (cf. courbe rouge Figure 2.1(a)), i.e.

$$p|k \backsim \beta(\alpha + k, \alpha + 2N - k)$$

La connaissance de loi a posteriori du paramètre θ permet de produire des estimateurs bayésiens ; les calculs du mode et/ou de la moyenne de la distribution a posteriori fournissent des estimations ponctuelles, l'intervalle qui contient 95% des plus grandes valeurs de la densité a posteriori [1] constitue une estimation ensembliste. Ainsi, un estimateur bayésien de la fréquence de l'allèle A est

$$\tilde{p} = \arg\max \Pr(p|k) = \frac{k+\alpha-1}{2N+2(\alpha-1)} = \frac{15}{38} \approx 0,39$$

valeur de p pour laquelle la densité a posteriori est maximale (mode, croix rouges sur la Figure 2.1).

L'estimation bayésienne est un compromis entre l'a priori et les observations (illustrations formule ci-dessus et Figure 2.1). Plus il y a de données, plus l'estimation bayésienne est proche de celle du fréquentiste (i.e. proche des observations). Inversemment, les estimateurs sont proches de l'a priori lorsque peu de données sont disponibles. Ces propriétés sont illustrées pour l'estimation de p sur la Figure 2.1(b). Lorsque le nombre d'individus échantillonnés augmente (resp. diminue), l'estimation a posteriori tend vers l'estimation fréquentiste (resp. le modèle Beta a priori).

La loi a priori apporte de l'information lorsque peu de données sont disponibles. Cependant le choix de la valeur du paramètre dont dépend la loi a priori est contestable. En effet, supposer que le niveau de différentiation génétique par rapport à la population ancestrale égale 20 ($\alpha = 10$) est subjectif. La démarche bayésienne conduit alors à introduire un a priori sur la loi a priori (hyper-prior) dans un modèle hiérarchique pour estimer α (hyper-paramètre). Par exemple, la distrbution a priori du paramètre α pourrait suivre une loi log-normale,

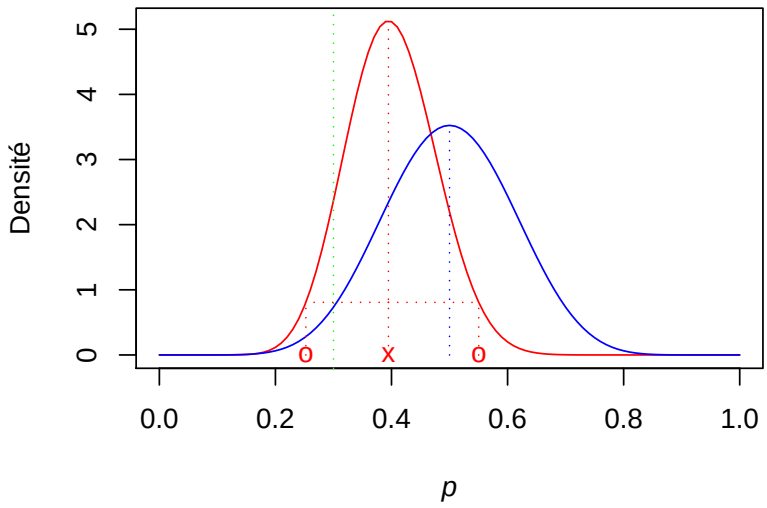
$$\log\alpha \backsim \mathcal{N}(0,1), \quad \text{i.e.} \quad \Pr(\alpha) = \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{(\log\alpha)^2}{2}\right) \tag{2.3}$$

2.2 Modèles bayésiens hiérarchiques

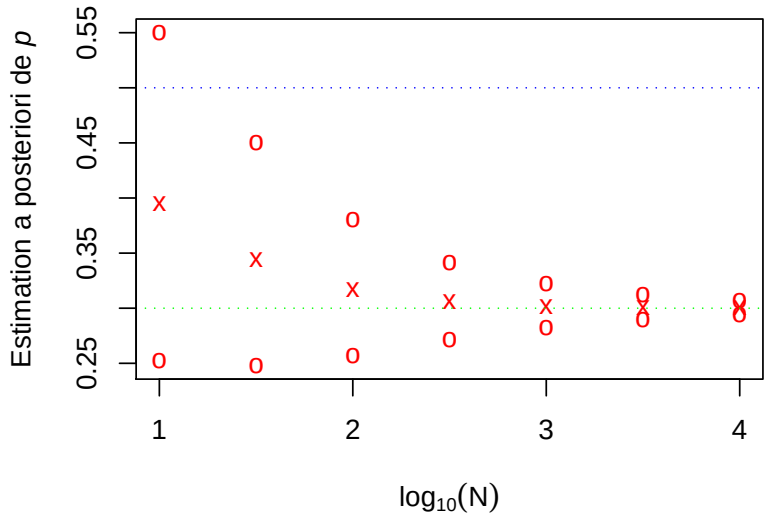
En statistique bayésienne, souvent, la loi a priori du paramètre θ dépend d'un autre paramètre, ψ, qui peut, lui aussi, avoir une loi a priori. Le modèle probabiliste contient alors une relation stochastique supplémentaire : la loi des données dépend du paramètre θ dont la loi dépend du paramètre ψ de loi déterminée. Ainsi, la réécriture de la formule de Bayes met en évidence une hiérarchie des paramètres dans le modèle probabiliste

$$\Pr(\theta,\psi|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\theta)\Pr(\theta|\psi)\Pr(\psi)}{\Pr(\mathcal{D})} \tag{2.4}$$

1. Highest Posterior Density Interval (HPDI).



(a) Distributions a priori et a posteriori de p.



(b) Influence de la quantité de données sur les estimateurs bayésiens (mode et HPDI).

FIGURE 2.1 – Inférence bayésienne de la fréquence p d'un allèle. En bleu l'a priori, en rouge l'a posteriori, en vert les données.

avec

$$\Pr(\mathcal{D}) = \int \Pr(\mathcal{D}|\theta)\Pr(\theta|\psi)\Pr(\psi)\,d\theta d\psi \tag{2.5}$$

L'*approche bayésienne hiérarchique* consiste à décrire des modèles probabilistes dont les paramètres s'organisent en couches successives par le biais de relations stochastiques interdépendantes. L'avantage de cette démarche est qu'elle autorise l'apport d'informations a priori sous la forme de modèles bien établis. Pour cette raison, les méthodes bayésiennes sont de plus en plus utilisées dans le domaine de la génétique des populations.

Les modèles bayésiens hiérarchiques se représentent par des graphes acycliques orientés [2] qui mettent en évidence les liens entre paramètres (cf. Figure 2.2). Les nœuds du graphe correspondent aux données (carrés) et aux paramètres (cercles), les arcs aux liens stochastiques du modèle hiérarchique. L'orientation des arcs donne la direction de la relation entre deux nœuds et permet de retrouver la formulation bayésienne (2.4).

L'objectif de la démarche est de déterminer la loi a posteriori des paramètres θ et ψ. Dans diverses situations, le calcul de l'expression de la loi a posteriori $\Pr(\theta, \psi|\mathcal{D})$ s'avère problématique. Dans l'exemple de la fréquence de l'allèle A, avec un hyper-prior de loi log-normale, le dénominateur de la loi a posteriori implique une intégrale généralisée. En effet, quelques lignes de calculs prouvent que

$$\Pr(k) = \int_0^{+\infty} \frac{\binom{2N}{k}}{\alpha\sqrt{2\pi}} \exp\left(-\frac{(\log\alpha)^2}{2}\right) \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \frac{\Gamma(\alpha+k)\Gamma(\alpha+2N-k)}{\Gamma(2\alpha+2N)}\,d\alpha \tag{2.6}$$

quantité délicate à évaluer

D'une part, la complexité des modèles et/ou de l'espace des paramètres complique le calcul du dénominateur $\Pr(\mathcal{D})$ de la loi a posteriori (cf. équations (2.4) et (2.5)). D'autre part, il n'est pas toujours possible de produire l'expression de la vraisemblance $\Pr(\mathcal{D}|\theta)$. L'utilisation de méthodes numériques permet de contourner ces problèmes et d'approcher la distribution a posteriori.

2.3 Algorithmes d'estimation

Lorsque l'expression de la loi a posteriori est trop difficile à déterminer, des algorithmes de simulation permettent d'approcher la distribution a posteriori. Par la suite plusieurs techniques vont être présentées, leur objectif est de simuler un échantillon de la loi a posteriori pour estimer les paramètres du modèle bayésien.

2. Directed Acyclic Graph (DAG).

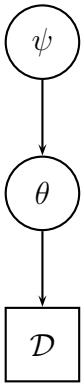


FIGURE 2.2 – Représentation graphique d'un modèle bayésien hiérarchique par un graphe acyclique orienté. Les nœuds carrés représentent les données, les nœuds circulaires les paramètres du modèle. Les arcs correspondent aux dépendances stochastique, la loi de $\mathcal{D}$ dépend de θ dont la loi dépend de ψ de loi déterminée.

2.3.1 Méthodes de Monte Carlo par chaîne de Markov (MCMC)

Les méthodes MCMC [3] simulent une chaîne de Markov dont la loi stationnaire est la loi souhaitée ($\Pr(\theta|\mathcal{D})$) [4], leur construction ne pose pas de grandes difficultés, contrairement à leur mise en pratique.

Les méthodes MCMC sont des procédés de simulations stochastiques très puissants qui permettent de contourner le (délicat) calcul de la constante $\Pr(\mathcal{D})$ (cf. équation (2.5)). La connaissance de la vraisemblance $\Pr(\mathcal{D}|\theta)$ et des loi a priori $\Pr(\theta)$ suffit pour implémenter ces méthodes, la loi a posteriori se calcule alors à une constante multiplicative près, la formulation bayésienne s'écrit

$$\Pr(\theta|\mathcal{D}) \propto \Pr(\mathcal{D}|\theta)\Pr(\theta) \tag{2.7}$$

Une chaîne de Markov est un processus aléatoire discret, $\{\theta_t; t \in \mathbb{N}\}$, dont la loi de l'état θ_t est conditionnée par le dernier état θ_{t-1} du processus. Les techniques MCMC reposent donc sur la construction (et la simulation) d'un noyau de transition qui décrit les lois conditionnelles d'une chaîne de Markov dont la distribution stationnaire suit la loi souhaitée.

La convergence de la chaîne de Markov vers une loi stationnaire est garantie lorsque la chaîne est *ergodique*, i.e. *irréductible* [5] et *apériodique* [6]. Par ailleurs, sous ces conditions d'ergodicité, pour générer la bonne loi stationnaire le noyau de transition doit être *réversible* [7].

L'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs permettent de simuler la loi a posteriori $\Pr(\theta|\mathcal{D})$ pour estimer les paramètres d'un modèle donné. Lorsque l'inférence consiste à comparer des modèles, les probabilités a posteriori de ces derniers sont évaluées grâce aux méthodes MCMC à saut réversible.

L'algorithme de Metropolis-Hastings (MH)

L'algorithme de Metropolis-Hastings (Metropolis et al. 1953, Hastings 1970) propose à chaque itération un nouvel état qui est ensuite accepté ou rejeté. Le paragraphe ci-dessous décrit le passage de θ_t à θ_{t+1}. L'algorithme MH permet de mettre à jour les paramètres dont la loi conditionnelle est difficile à simuler.

Soit $q(.|\theta)$ une distribution qui permet de proposer un état pour la chaîne de Markov à partir de l'état actuel, θ. L'état candidat généré, θ', sera le nouvel état

3. Markov Chain Monte Carlo.

4. Pour ne pas surcharger les notations paramètres et hyper-paramètres sont confondus.

5. Tout état de la chaîne de Markov est atteignable en un temps fini quelque soit l'état initial.

6. La chaîne de Markov ne boucle pas sur un ensemble fini d'états.

7. La probabilité d'être dans un état pour passer dans un autre est égale à la probabilité de faire le chemin inverse.

du processus avec la probabilité

$$\alpha(\theta, \theta') = \min\left(1, \frac{\Pr(\theta'|\mathcal{D})q(\theta|\theta')}{\Pr(\theta|\mathcal{D})q(\theta'|\theta)}\right) \tag{2.8}$$

sinon la chaîne restera dans son état actuel.

Pour compléter le noyau de transition de la chaîne de Markov, il faut donner la probabilité pour que le processus reste dans le même état θ soit

$$1 - \int q(\theta'|\theta)\alpha(\theta, \theta')\, d\theta' \tag{2.9}$$

Le noyau de transition $\alpha(\theta, \theta')q(\theta'|\theta)$ est bien réversible (par construction) et, lorsque la propriété d'ergodicité est vérifiée, admet $\Pr(\theta|\mathcal{D})$ pour loi stationnaire. En effet, quelques calculs démontrent que la proporiété de réversibilité est vérifiée, i.e.

$$\Pr(\theta|\mathcal{D})\alpha(\theta, \theta')q(\theta'|\theta) = \Pr(\theta'|\mathcal{D})\alpha(\theta', \theta)q(\theta|\theta') \tag{2.10}$$

Le problème du calcul de $\Pr(\mathcal{D})$ disparait et justifie la formulation bayésienne (2.7). En effet, la probabilité d'accepter le nouvel état, $\alpha(\theta, \theta')$, fait intervenir le ratio des distributions a posteriori, d'après la formule de Bayes,

$$\frac{\Pr(\theta'|\mathcal{D})}{\Pr(\theta|\mathcal{D})} = \frac{\Pr(\mathcal{D}|\theta')\Pr(\theta')/\Pr(\mathcal{D})}{\Pr(\mathcal{D}|\theta)\Pr(\theta)/\Pr(\mathcal{D})} = \frac{\Pr(\mathcal{D}|\theta')\Pr(\theta')}{\Pr(\mathcal{D}|\theta)\Pr(\theta)}$$

Ainsi l'algorithme ne nécessite la connaissance de $\Pr(\theta|\mathcal{D})$ qu'à une constante multiplicative près.

L'algorithme 1 permet de simuler une chaîne de Markov dont la loi stationnaire est la loi a posteriori $\Pr(\theta|\mathcal{D})$. Le comportement de la chaîne s'intuite à partir de l'équation (2.8) : à l'équilibre, l'exploration de l'espace des paramètres se concentre sur les régions pour lesquelles la probabilité (ou densité) a posteriori est forte. En effet, les sauts vers des régions de plus haute probabilité a posteriori sont presque toujours acceptés, les petits (resp. grands) sauts vers des densités un peu (resp. beaucoup) plus faibles sont parfois (resp. rarement) acceptés (illustration Figure 2.3).

Echantillonneur de Gibbs

Lorsque la loi conditionnelle de certains paramètres est connue et facile à simuler, l'échantilloneur de Gibbs génère directement un échantillon de la loi conditionnelle en question. Il s'agit en fait d'un cas particulier de l'algorithme MH dans lequel la distribution $q(.|\theta)$ est égale à la loi conditionnelle de θ. Du fait de cette utilisation toutes les mises à jour des paramètres concernés sont acceptées car, dans ce cas, la probabilité de passer de l'état θ à l'état θ', $\alpha(\theta, \theta')$, vaut un.

Algorithme 1 Algorithme de Metropolis-Hastings

{Initialiser la chaîne de Markov}
$\theta_0 \backsim \Pr(\theta)$
répéter
  {Proposer un nouvel état}
  $\theta' \backsim q(.|\theta_t)$
  {Calculer la probabilité d'accepter ce nouvel état}
  $\alpha(\theta, \theta') = \min\left(1, \frac{\Pr(\mathcal{D}|\theta')\Pr(\theta')q(\theta|\theta')}{\Pr(\mathcal{D}|\theta)\Pr(\theta)q(\theta'|\theta)}\right)$
  {Accepter ou rejeter la mise à jour avec la probabilité α}
  $u \backsim \mathcal{U}(0,1)$
  si $u \leq \alpha(\theta_t, \theta')$ **alors**
    $\theta_{t+1} = \theta'$
  sinon
    $\theta_{t+1} = \theta_t$
  finsi
  {Passer à l'itération suivante}
  $t = t + 1$
jusqu'à l'atteinte du régime stationnaire et l'obtention d'un échantillon de la loi a posteriori

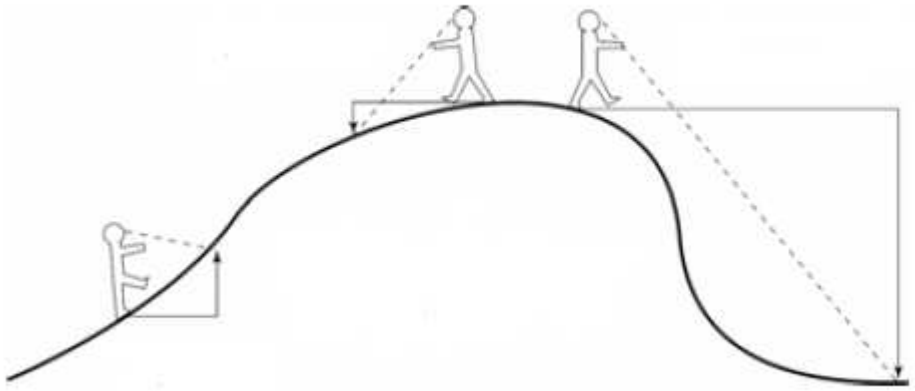


FIGURE 2.3 – Comportement de l'algorithme de Metropolis-Hastings. Le robot se promène plus souvent aux alentours du sommet et descend rarement au fond des vallées.

Méthodes MCMC à saut réversible (RJMCMC)

Les méthodes MCMC à saut réversible (Green 1995) étendent les possibilités des techniques MCMC à la sélection de modèles (dont le nombre de paramètres peut varier). Le but des méthodes RJMCMC [8] est d'estimer les probabilités a posteriori des modèles considérés pour identifier celui(eux) qui explique(nt) le mieux les données observées et pondérer les estimations des paramètres.

Chaque modèle $\mathcal{M}$ comprend un jeu de paramètre $\theta_{\mathcal{M}}$ dont la dimension $n_{\mathcal{M}}$ peut changer d'un modèle à l'autre. Si $\Pr(\mathcal{M})$ est la probabilité a priori du modèle $\mathcal{M}$, la formule de Bayes donne la distribution a posteriori

$$\Pr(\theta_{\mathcal{M}}, \mathcal{M}|\mathcal{D}) \propto \Pr_{\mathcal{M}}(\mathcal{D}|\theta_{\mathcal{M}})\Pr_{\mathcal{M}}(\theta_{\mathcal{M}})\Pr(\mathcal{M}) \tag{2.11}$$

avec $\Pr_{\mathcal{M}}(\mathcal{D}|\theta_{\mathcal{M}})$, la vraisemblance des données selon le modèle $\mathcal{M}$ et $\Pr_{\mathcal{M}}(\theta_{\mathcal{M}})$ la distribution a priori des paramètres selon le même modèle.

L'algorithme RJMCMC, de façon analogue à l'algorithme MH, propose une mise à jour du modèle. Si le modèle courant est $\mathcal{M}$, le modèle $\mathcal{M}'$ est proposé avec la probabilité $\Pr(\mathcal{M} \to \mathcal{M}')$. Il s'agit ensuite de passer de $\theta_{\mathcal{M}}$ à $\theta_{\mathcal{M}'}$ sans oublier de prendre en compte les changements de dimension. Dans ce but, un vecteur u de taille d est généré à partir d'une distribution $q_{\mathcal{M},\mathcal{M}'}$ indépendante de $\theta_{\mathcal{M}}$. Les paramètres du nouveau modèle se calculent à partir du difféomorphisme [9] $g_{\mathcal{M},\mathcal{M}'}$,

$$(\theta'_{\mathcal{M}}, u') = g_{\mathcal{M},\mathcal{M}'}(\theta_{\mathcal{M}}, u) \tag{2.12}$$

de telle sorte que les dimensions des espaces de départ et d'arrivée de $g_{\mathcal{M},\mathcal{M}'}$ correspondent, i.e.

$$n'_{\mathcal{M}} + d' = n_{\mathcal{M}} + d \tag{2.13}$$

Le passage de $(\theta_{\mathcal{M}}, \mathcal{M})$ à $(\theta_{\mathcal{M}'}, \mathcal{M}')$ est accepté avec la probabilité

$$\alpha(\mathcal{M}, \mathcal{M}') = \min\left(1, \frac{\Pr(\theta_{\mathcal{M}'}, \mathcal{M}'|\mathcal{D})q_{\mathcal{M}',\mathcal{M}}(u')\Pr(\mathcal{M}' \to \mathcal{M})}{\Pr(\theta_{\mathcal{M}}, \mathcal{M}|\mathcal{D})q_{\mathcal{M},\mathcal{M}'}(u)\Pr(\mathcal{M} \to \mathcal{M}')} \left|\frac{\partial g_{\mathcal{M},\mathcal{M}'}(\theta_{\mathcal{M}}, u)}{\partial(\theta_{\mathcal{M}}, u)}\right|\right) \tag{2.14}$$

Les probabilités a posteriori des modèles se calculent à partir des fréquences de visite de chaque modèle. Ces probabilités permettent de tenir compte de l'incertitude liée au choix du modèle lors de l'estimation des paramètres.

Dans l'exemple de l'estimation de la fréquence de l'allèle A, deux modèles ont été décrits : dans le premier, le niveau de différentiation génétique par rapport à la population ancestrale est fixé ($\alpha = 10$, faible) alors que dans le second il est aléatoire (hyper-paramètre $\log \alpha \backsim \mathcal{N}(0, 1)$, fort). Les simulations RJMCMC permettent de comparer ces deux modèles dans une seule et même chaîne.

8. Reversible Jump Markov Chain Monte Carlo.
9. Fonction différentiable bijective dont la réciproque est également différentiable.

Implémentation

La mise en pratique des méthodes MCMC décrites dans ce chapitre présente des difficultés qui peuvent compliquer l'interprétation des résultats. En particulier, le temps d'atteinte du régime stationnaire est inconnu, il convient d'attendre un certain nombre d'itérations (*burn-in*) avant d'échantillonner la chaîne de Markov. La stratégie d'exploration de l'espace des paramètres, i.e. le choix de $q(.|\theta)$ et des conditions initiales, influencent le temps de convergence. La conception d'un noyau de transition adapté au problème posé est donc essentielle à la qualité des estimations a posteriori. Par ailleurs, il est difficile d'obtenir un échantillon de réalisations indépendantes de la loi stationnaire. En effet, les valeurs issues de la simulation d'une chaîne de Markov sont généralement corrélées. Il convient alors d'observer un certain intervalle de temps (*thining*) entre deux échantillonnages de la chaîne.

L'estimation de la fréquence de l'allèle A permet d'illustrer quelques uns des problèmes liés à l'implémentation des méthodes MCMC (Figure 2.4). D'après la formule de Bayes, la distribution a posteriori des paramètres p et α est connue à une constante multiplicative près, i.e.

$$\Pr(p, \alpha|k) \propto p^{\alpha+k-1}(1-p)^{\alpha+2N-k-1}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha+1)}\alpha^{-1}\exp\left(-\frac{(\log\alpha)^2}{2}\right) \quad (2.15)$$

Le membre de droite de la relation ci-dessus est une fonction des deux paramètres du modèle dont les courbes de niveau sont représentées sur les Figures 2.4(a) (à gauche) et 2.4(b). Les régions de l'espace des paramètres de forte (resp. faible) densité a posteriori sont représentées en magenta (resp. cyan). Les simulations MCMC explorent la paysage ainsi formé. Les mises à jour des paramètres p et α sont détaillées dans l'appendice ? ?. Les simulations comprennent 3000 itérations, les 500 premières sont représentées en rouge, les 2500 suivante en bleu.

La figure 2.4(a) met en évidence le temps d'atteinte de la loi stationnaire. Après 500 itérations la chaîne de Markov n'a pas encore visité tout l'espace et n'a donc pas convergé vers la loi a posteriori. Les histogrammes obtenus à partir de cet échantillonnage sont peu représentatifs de la loi a posteriori. Les 2500 itérations qui suivent complètent l'exploration et permettent d'obtenir des histogrammes de meilleure qualité.

La figure 2.4(b) illustre trois stratégies d'exploration de l'espace des paramètres. Celle de gauche consiste à proposer des mises à jour de faible amplitude dans l'espace des paramètres ; dans ce cas presque toutes les transitions sont acceptées mais l'exploration de l'espace des paramètres est lente. Celle de droite propose de faire des sauts sur de grandes distances ; cette méthode conduit à un grand nombre de rejets et à un échantillonnage grossier. La stratégie intermédiaire (au centre) est celle recherchée ; les taux de rejets obtenus conduisent à une meilleure exploration

de l'espace des paramètres.

2.3.2 Méthodes bayésiennes approchées (ABC)

Les méthodes ABC[10] sont particulièrement prisées dans les cas où le calcul de la vraisemblance $\Pr(\mathcal{D}|\theta)$ est trop coûteux ou indéterminé. Le principe de ces techniques repose sur la simulation de données à partir du modèle bayésien dont les paramètres sont générés selon leur loi a priori. L'idée est que les données simulées « suffisament proches » des données observées permettent d'approcher la loi a posteriori. Ici « suffisament proches » signifie que la distance ρ entre les statistiques S et S' qui décrivent données réelles, $\mathcal{D}$, et simulées, $\mathcal{D}'$, est inférieure à un seuil donné, $\epsilon > 0$.

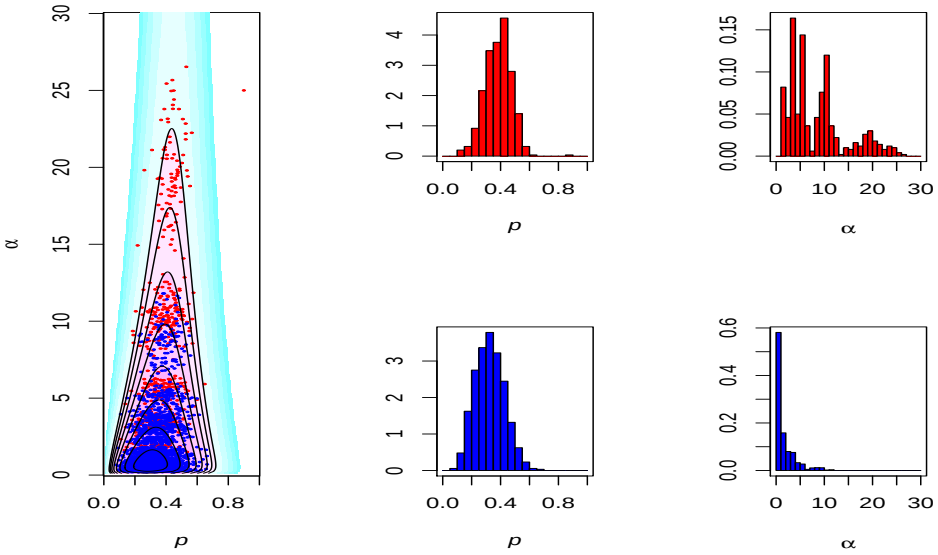
L'algorithme 2 génére un échantillon de n paires (θ_i, s_i) indépendantes qui permet d'estimer la loi a posteriori de façon approximative. Il s'agit en fait d'une méthode de rejet qui simule la distribution $\Pr(\theta|\rho(S, S') < \epsilon)$ pour approcher $\Pr(\theta|S)$ lorsque la valeur de ϵ tend vers zéro. Ainsi les valeurs θ_i du paramètre pour lesquelles la distance $\rho(s, s_i)$ est faible constitue un échantillon approximatif de la loi a posteriori.

Deux modifications apportées par Beaumont et al. (2002) corrigent certaines approximations pour produire les estimations. La première amélioration consiste à pondérer les valeurs θ_i du paramètre selon la distance $\rho(s, s_i)$. Plus les statistiques descriptives des données simulées sont proches de celles des données observées, moins il y a d'incertitude sur la valeur du paramètre à estimer. L'autre amélioration consiste à ajuster les valeurs du paramètres par une régression linéaire locale pour corriger les écarts entre s et s_i. L'utilisation d'une telle régression requiert les hypothèses de linéarité et d'homoscedaticité[11] au voisinnage de s.

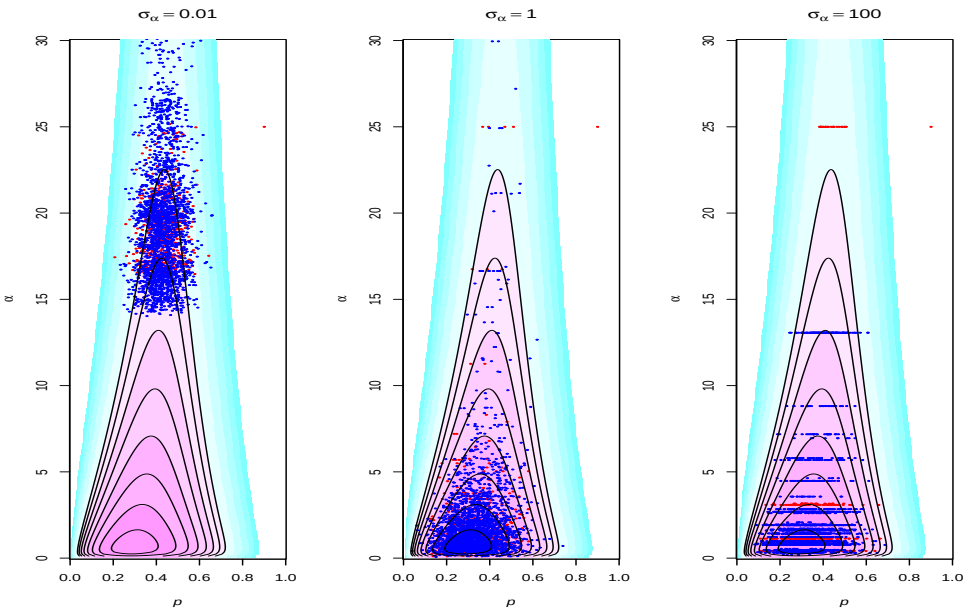
L'implémentation des méthodes ABC repose sur des algorithmes de simulation de modèles probabilistes. La génération des données simulées est parallélisable et les échantillons obtenus sont indépendants. La principale difficulté réside dans le choix des statistiques descriptives adaptées au problème étudié et du seuil de tolérance ϵ. En effet, une statistique donnée peut améliorer les estimations d'un paramètre au détriment des estimations d'un autre. Le seuil de rejet a des effets opposés sur le biais et la variance des estimations : augmenter la tolérance réduit la variance (régression de meilleure qualité) mais augmente le biais (violation des hypothèses de linéarité et d'homoscedaticité).

L'approche bayésienne hiérarchique et les méthodes MCMC et ABC associées constituent de puissants outils pour la génétique des populations. Leur popularité dans ce domaine s'expliquent par leurs aspects flexibles (modélisation) et pratiques

10. Approximate Bayesian Computation
11. Constance de la variance des résidus



(a) Temps d'atteinte de la loi stationnaire.



(b) Stratégie d'exploration.

FIGURE 2.4 – Implémentation des méthodes MCMC pour l'estimation de la fréquence de l'allèle A.

Algorithme 2 Algorithme ABC

$i = 0$
{Calculer les statistiques descriptives pour les données observées}
$s = S(\mathcal{D})$
répéter
  {Générer les paramètres du modèle selon la loi a priori}
  $\theta \backsim \Pr(\theta)$
  {Simuler les données selon le modèle}
  $\mathcal{D}' \dashleftarrow \theta$
  {Calculer les statistiques descriptives pour les données simulées}
  $s' = S(\mathcal{D}')$
  {Accepter les paramètres selon le seuil de tolérance pour les distances}
  si $\rho(s, s') < \epsilon$ **alors**
    $\theta_{i+1} = \theta$
    $i = i + 1$
  finsi
jusqu'à obtenir un échantillon de taille souhaitée

(estimation) qui permettent de les appliquer en génétique des populations.

Les méthodes présentées dans la partie II combinent algorithme MH, échantillonneur de Gibbs et techniques RJMCMC pour estimer l'effet des variables environnementales sur les flux de gènes dans un cadre bayésien. La méthode décrite dans la partie III utilise l'approche ABC pour estimer l'influence des variables environnementales sur la sélection naturelle.

Deuxième partie

L'estimation des taux de migration

Chapitre 3

Migration, flux de gènes et structure spatiale des populations

L'étude des flux de gènes est fondamentale en écologie, en génétique des populations, en biologie de la conservation et pour la gestion des espèces sauvages. La question essentielle et commune à ces applications est de prédire et de mesurer l'importance des processus de dispersion dans le maintien de la biodiversité. Il s'agit de prévoir l'impact de la fragmentation de l'habitat ou de changements environnementaux sur la distribution spatiale de la diversité génétique. Par ailleurs, l'estimation de l'intensité des flux géniques permet de quantifier le niveau d'interdépendance des populations et d'établir des plans de conservation adaptés.

3.1 Les processus de dispersion

Les processus de dispersion jouent un rôle central dans la structuration spatiale de la diversité génétique. La colonisation de nouveaux territoires ou les migrations entre populations assurent la diffusion des allèles sur l'aire de répartition des espèces. Les échanges de gènes entre populations modifient la structure génétique des populations et diminuent les variations spatiales des fréquences alléliques. Les schémas de dispersion dépendent de l'espèce, de son habitat et du succès reproductif des nouveaux arrivants.

Les modes de dispersion varient selon les espèces, ils peuvent être actifs ou passifs. Par exemple, les oiseaux ou les mammifères se déplacent par eux-mêmes alors que le pollen ou les graines des plantes sont transportés par les vents ou par les insectes. Par ailleurs, la dispersion peut intervenir à différents stades du cycle de vie, elle peut concerner les gamètes (e.g. le pollen chez les plantes), les individus ou des groupes d'individus (e.g. oiseaux ou poissons migrateurs). Enfin les distances de dispersion dépendent également de l'espèce considérée.

La distribution géographique des espèces est influencée par l'apparition ou la disparition d'obstacles à la dispersion. Les chaînes de montagnes, les océans ou d'autres facteurs environnementaux peuvent constituer des barrières géographiques responsables de l'isolement des populations. Les contraintes de l'habitat peuvent ainsi interrompre momentanément ou durablement les échanges de matériel génétique et conduire à l'apparition de nouvelles espèces.

Les modifications des fréquences allèliques s'opèrent lors de la reproduction entre occupants et nouveaux arrivants. Le régime de reproduction influence donc le devenir des gènes qui ont dispersé.

Le processus de dispersion étudié dans cette partie est la migration. Du point de vue de l'évolution les populations correspondent à des paquets de gènes (sous-populations) et la migration à des flux de gènes (ou flux géniques). L'intensité de ces flux, mesurée par les taux de migration, dépend donc de la mobilité de l'espèce, des contraintes environnementales et du succès reproductif des migrants.

3.2 Les modèles de la génétique des populations

Sur le plan théorique, la génétique des populations permet de prédire l'évolution des populations selon l'intensité des flux géniques. Les modèles mathématiques développés (entre autres) par Wright, Haldane, Fisher, Kimura, Slatkin et Kimura ont permis d'expliquer le rôle des processus de dispersion par rapport aux autres forces évolutives. Le paragraphe qui suit présente les principaux résultats de ces travaux théoriques sur la migration.

Le rôle des flux de gènes en tant que force évolutive est de s'opposer à la différentiation génétique [1] due à la dérive génétique et à la sélection naturelle. En effet, en l'absence de migration, l'évolution des fréquences alléliques est liée aux effectifs finis des sous-populations et à l'adaptation aux conditions environnemetales locales. La résultante de ces deux forces conduit donc à la différentiation génétique des sous-populations alors que l'effet de la migration est de redistribuer la diversité génétique. L'autre rôle des flux géniques est la diffusion des mutations favorables apparues dans une sous-population et le maintien du potentiel adaptatif des espèces.

3.2.1 Modèles de migration

La littérature de la génétique des populations est riche en modèles de migration dans une population subdivisée. Ce paragraphe passe en revue quelques uns des modèles les plus couramment rencontrés (cf. Figure 3.1). Les changements de la

1. Divergence des structures génétiques des sous-populations.

fréquence p d'un gène à locus biallélique [2] permettent de mieux comprendre l'effet de la migration sur l'évolution des fréquences alléliques.

Modèle continent-îles

Le modèle continent-îles [3] constitue l'exemple le plus simple de schéma migratoire. Une population est subdivisée en plusieurs sous-populations (les îles) : un continent dont la composition génétique ne change pas est entouré de plusieurs petites îles de même effectif N qui reçoivent à chaque génération une proportion m de migrants du continent (cf. Figure 3.1(a)). Selon ce scénario, lors de la prochaine génération la fréquence de l'allèle sur l'île i sera

$$p'_i = (1 - m)p_i + m\bar{p} \tag{3.1}$$

où $\bar{p}$ est la fréquence de l'allèle sur le continent. La situation d'équilibre sera atteinte lorsque la fréquence de l'allèle sur chaque l'île sera la même que sur le continent.

Modèle en îles de Wright

Le modèle en îles de Wright [4] s'inspire d'un archipel (cf. Figure 3.1(b)). Bien que peu réaliste ce modèle explique l'effet de l'interaction entre dérive génétique et migration sur la différentiation génétique (cf. §3.2.2).

Une population est subdivisée en I îles interconnectées et de même effectif constant N. Chaque île reçoit à chaque génération le même nombre de migrants de chacunes des autres sous-populations. Le taux de migration par génération, noté m, est également constant. Selon ce schéma, la fréquence de l'allèle dans la sous-population i sera

$$p'_i = (1 - m)p_i + \frac{m}{I - 1}\sum_{j \neq 1} p_j \tag{3.2}$$

lors de la prochaine génération. Quelque calculs permettent de montrer que la situation d'équilibre sera atteinte lorsque la fréquence de l'allèle sera la même sur toutes les îles, soit la fréquence de l'allèle $\bar{p}$ dans la population totale (qui reste constante).

Modèles en treillis de Kimura

Kimura a introduit les modèles en treillis [5] par analogie avec les dalles d'un jardin japonais. Les modèles en treillis sont plus réalistes que les modèles en îles

2. Par la suite p_i désignera la fréquence de l'alléle dans la sous-population i.
3. Mailand-island model.
4. Wright's island model.
5. Stepping-stone models.

dans la mesure où ils prennent en compte une structure spatiale.

Une population est subdivisée en plusieurs sous-populations qui n'échangent des migrants qu'avec les sous-populations voisines. Les sous-populations sont distribuées régulièrement le long d'un habitat uni ou bidimensionnel, elles possèdent toutes le même nombre de voisins k ($k = 2$ en 1D, $k = 4$ en 2D, Figures 3.1(c) et 3.1(d)). Les échanges de migrants entre sous-populations sont symétriques et s'effectuent avec le même taux de migration $\frac{m}{k}$. La fréquence de l'allèle dans la sous-population i lors de la génération suivante sera alors

$$p_i' = \frac{m}{k} \sum_{j_i} p_{j_i} + (1 - m) p_i \tag{3.3}$$

où les indices j_i désignent les sous-populations voisines de la sous-populations i.

Isolement par la distance

Les modèles d'isolement par la distance constituent une extension des modèles en treillis dans lesquels la dispersion est limitée dans l'espace. L'idée de base est que deux individus ont plus de chance de se reproduire s'ils sont géographiquement. La migration est fonction du noyau de dispersion : une majorité d'individus dispersent sur de courtes distances. Ainsi les taux de migration diminuent avec la distance entre sous-populations, et la probabilité de migrer est inversemment proportionnelle aux distances entre sous-populations, i.e.

$$\Pr(i \to j) \propto \frac{1}{d_{ij}} \tag{3.4}$$

où d_{ij} est la distance géographique entre les sous-populations i et j.

Modèle matriciel

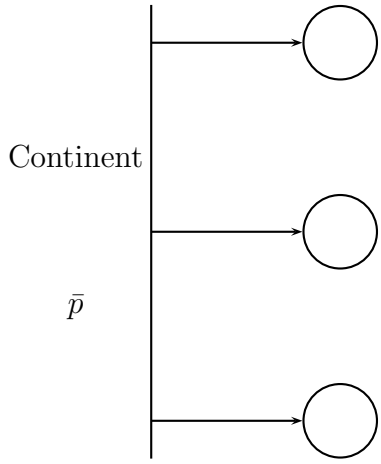
Le modèle matriciel généralise les modèles présentés ci-dessus, la migration n'est pas forcément symétrique et diffère selon les paires de sous-populations (cf. Figure 3.1(e)). Une matrice de migration $\boldsymbol{M} = (m_{ij})$, où m_{ij} est la proportion d'individus de la population i qui proviennent de la sous-population j, est alors utilisée pour stockée les taux de migration. Dans ce cas, la fréquence de l'allèle dans la sous-population i sera

$$p_i' = \sum_j m_{ij} p_j \tag{3.5}$$

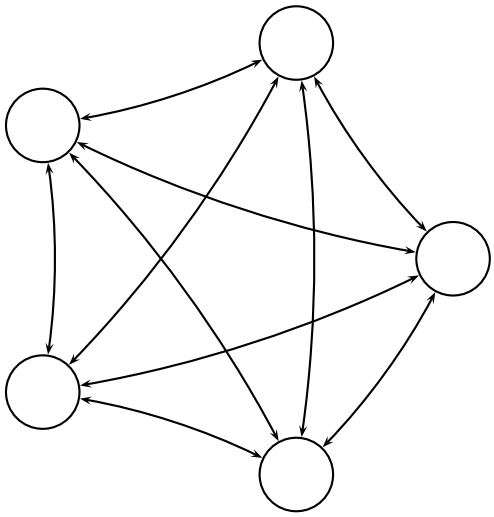
lors de la prochaine génération. La relation précédente se réécrit sous la forme matricielle

$$\boldsymbol{p}' = \boldsymbol{M}\boldsymbol{p} \tag{3.6}$$

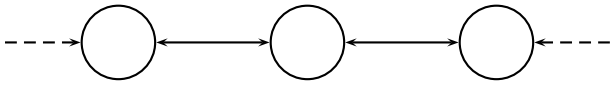
où $\boldsymbol{p} = (p_i)$ est le vecteur des fréquences alléliques dans chaque sous-population.



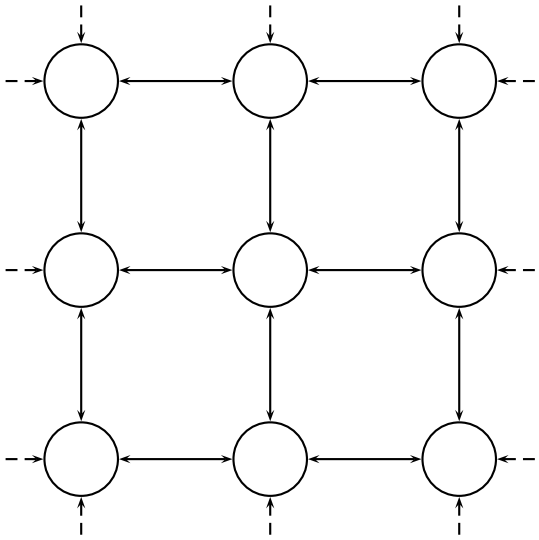
(a) Modèle continent-îles.



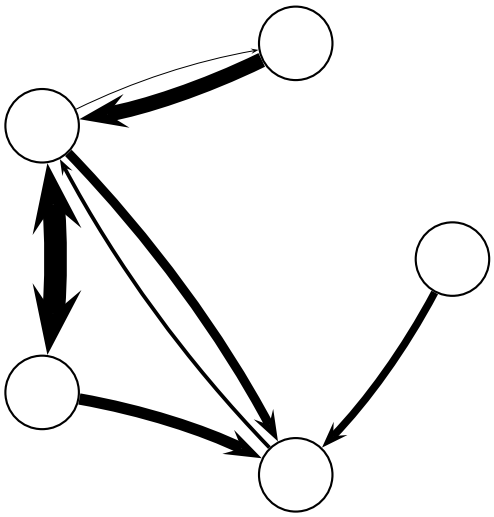
(b) Modèle en îles de Wright.



(c) Modèle en treillis 1D.



(d) Modèle en treillis 2D.



(e) Modèle avec taux de migration arbitraires.

FIGURE 3.1 – Modèles de migration dans une population subdivisée. Les cercles représentent les sous-populations et les arcs les flux de gènes. L'épaisseur des arcs est proportionnelle au taux de migration.

3.2.2 Mesure de la structure des populations

Lorsqu'une population est structurée en sous-populations, la différentiation génétique rend compte de la distribution de la variabilité génétique entre sous-populations par rapport à la population totale. S'il y a peu de différences dans les structures alléliques des sous-populations, le niveau de différentiation génétique est faible. Dans le cas inverse, les sous-populations sont fortement différentiées.

Wright (1951) a introduit un indice, F_{ST}, qui mesure la corrélation entre variabilité génétique entre sous-populations et structure. Lorsque le coefficient F_{ST} est proche de 0 (resp. 1), la structure est faible (resp. forte). Pour un locus biallélique, le niveau de différentiation génétique est défini par la relation

$$F_{ST} = \frac{\sigma^2}{p(1-p)} \tag{3.7}$$

avec p la fréquence de l'un des deux allèles dans la population totale et σ^2 la variance de cette fréquence entre sous-populations.

Cas du modèle en îles de Wright L'un des schémas migratoires les plus étudiés est le modèle en îles de Wright. Pour un locus neutre (i.e. non soumis à la sélection) et en l'absence de mutations, Wright a montré qu'à l'équilibre entre dérive et migration

$$F_{ST} \approx \frac{1}{1+4Nm} \tag{3.8}$$

où Nm est le nombre de migrants qui participent à la reproduction dans chaque sous-population à chaque génération. La relation (3.8) ci-dessus permet d'établir que la dérive différentie considérablement les sous-populations ($F_{ST} > 0,2$) lorsqu'il y a moins d'un migrant par génération ($Nm < 1$).

Des modèles de migration complexes qui impliquent plusieurs forces évolutives ou des scénarios démographiques différents permettent de prévoir l'évolution de la structure spatiale de la diversité génétique. En particulier, les modèles de métapopulations ont été introduits pour mieux tenir compte de la dynamique des populations. Le développement des moyens de calcul facilite la simulation de modèles de migration élaborés pour valider les résultats théoriques. Cependant l'estimation des taux de migration dans les populations naturelles reste un problème difficile.

3.3 L'estimation des taux de migration

La mesure des taux de migration est, par nature, problématique ; elle requiert des connaissances sur la mobilité de l'espèce, sur les contraintes de son habitat,

et sur le succès reproductif des migrants (cf. 3.1). Deux types d'approches ont été développés pour estimer l'intensité des flux géniques dans les populations naturelles. Les méthodes *directes* cherchent à évaluer les distances de dispersion et le succès reproductif à partir du suivi des individus sur le terrain. Les méthodes *indirectes* utilisent l'information génétique pour estimer les paramètres de modèles de migration.

Méthodes directes

Les méthodes directes utilisent des approches purement écologiques qui consistent à observer directement les individus d'une espèce dans leur milieu naturel. Différentes méthodes permettent de suivre la dispersion : l'observation directe, les techniques de capture-marquage-recapture ou de radio-localisation. De manière générale l'approche directe donne des informations sur la mobilité des individus d'une espèce dans un habitat donné.

L'observation de la dispersion ne permet pas à elle seule d'estimer l'intensité des flux géniques, encore faut-il évaluer le succès reproductif des migrants. Ces derniers se déplacent pour trouver un habitat approprié pour se reproduire, mais cette recherche peut ne pas aboutir. Les cas de dispersion passive (e.g. transport de graines ou du pollen par le vent ou les insectes) illustre le fait que l'observation du succès reproductif est problématique.

Le suivi des populations in situ n'est possible que pour un nombre limité d'individus et d'espèces et la mise en œuvre des techniques directes est très coûteuse. L'étendue de l'aire géographique étudiée et/ou le mode de dispersion (e.g. gamétique) compliquent l'observation des déplacements. De plus, la durée du suivi des populations dépend du cycle de vie et donc du temps de renouvellement des générations de l'espèce considérée.

Les mesures directes des taux de migration constituent des instantanés (la durée des observations) des flux géniques mais ne rendent pas comptent de l'évolution des espèces. Par opposition, les méthodes indirectes détectent les traces de la migration à plus long terme par l'analyse de l'information génétique.

Méthodes indirectes

Les méthodes indirectes utilisent les modèles de la génétique des populations pour estimer les taux de migration à partir de l'information contenue dans les gènes. Ces techniques s'appuient sur la distribution spatiale de la diversité génétique observée à partir des fréquences alléliques, des séquences d'ADN, ou des génotypes multilocus. Trois types d'approches ont été développées selon le type de données utilisées.

L'avantage de ces approches par rapport aux méthodes directes est qu'elles permettent d'échantillonner un plus grand nombre d'invidus. Par ailleurs l'utilisation de l'information génétique intègre l'histoire évolutive de l'espèce étudiée. De ce point de vue les approches indirectes complètent les estimations obtenues à partir des méthode directes.

Approches basées sur les F_{ST} Les approches basées sur la statistique F_{ST} de Wright utilisent la relation (3.8) pour estimer le nombre de migrants par génération Nm d'un modèle en îles. Leur principe repose sur le calcul d'un estimateur de la différentiation génétique à partir des fréquences alléliques observées. D'autres indices que le F_{ST} peuvent être utilisés (e.g. θ, Weir and Cockerham 1984 ; N_{ST}, Lynch and Crease 1990 ; $< F_{ST} >$, Hudson et al. 1992 ; Φ_{ST}, Excoffier and Smouse 1994 ; ρ_{ST}, Rousset 1996, G_{ST}, Nei 1973 ; R_{ST} Slatkin 1995).

Ce type d'approches est critiqué car il suppose que les populations s'organisent selon un modèle en îles (Bossart and Prowell 1998 ; Whitlock and McCauley 1999). En effet, les hypothèses du modèle en île semblent peu réalistes et sont souvent invérifiées pour les populations naturelles (grand nombre de sous-population, taille et taux de migration constants, migration symétrique).

D'autres approches, plus récentes, utilisent l'information contenu dans les séquences d'ADN ou les génotypes multilocus pour estimer les paramètres de modèles plus réalistes. Deux approches complémentaires mesurent l'intensité des flux géniques sur des échelles de temps différentes. Les estimations des taux de migration issues du coalescent retracent les flux géniques sur le long terme alors que celles produites par les approches hors-équilibre les migrations récentes.

Approche coalescent Des méthodes d'estimation des taux de migration basées sur la théorie du coalescent (Kingman 1982a, 1982b) ont été developpées (e.g. GENTREE, Bahlo and Griffiths 2000 ; MDIV, Nielsen and Wakeley 2001 ; MIGRATE, Beerli and Felsenstein 2001). Les approches proposées utilisent les séquences d'ADN pour estimer la généalogie des gènes selon différents schémas. Les modèles utilisés autorisent des flux assymétriques et/ou des changements démographiques selon des scenarios proches de la réalité des populations naturelles.

Approche hors équilibre D'autres approches utilisent les déséquilibres gamétiques pour estimer les taux de migration à partir de génotypes multilocus. En effet, sur le court terme, l'empreinte des flux de gènes s'observe sur les génotypes de migrants ou de leur(s) descendant(s). Les modèles utilisés reposent sur peu d'hypothèses mais n'autorisent que l'estimation des taux de migration récents.

Les méthodes et les tests d'assignation permettent d'identifier les migrants (e.g. Paetkau et al. 1995 ; IMMANC, Rannala and Mountain 1997 ; STRUCTURE, Prit-

chard et al. 2000 ; BAPS, GENELAND, Guillot et al. 2005 ; Corander and Marttinen 2006 ; TESS François et al. 2006) et de retrouver les populations d'origine des individus. Cependant ces méthodes n'estiment pas directement les taux de migration.

La méthode implémentée dans le logiciel BAYESASS (Wilson and Rannala 2003) permet d'estimer directement les taux de migration à partir de génotypes multilocus.

3.4 Méthodes bayésiennes pour l'estimation des taux de migration

L'estimation des taux de migration est un problème fondamental pour l'étude de la structure génétique des populations. Deux méthodes bayésiennes ont été développées pour évaluer les flux de gènes récents à partir de génotypes multilocus (BAYESASS, Wilson and Rannala 2003 ; BIMR Faubet and Gaggiotti 2008). Ces deux approches reposent sur des modèles bayésiens hiérarchiques dont les paramètres sont estimés par les techniques MCMC (cf. Chapitre 2). La suite de cette partie est consacrée à leur étude.

Un premier article présente l'évaluation du logiciel BAYESASS à partir de donner simulées selon différents scénarios. Un deuxième article propose une nouvelle méthode bayésienne pour estimer l'influence des facteurs environnementaux sur la migration.

Chapitre 4

Article I

Ce chapitre présente l'article de Faubet et al. (2007) sur l'évaluation de la méthode de Wilson and Rannala (2003). Cette étude repose sur l'utilisation de simulations selon des modèles dont les paramètres sont contrôlés. Les données ainsi générées sont ensuite analysées avec BAYESASS pour mesurer les écarts entre les estimations a posteriori et les vraies valeurs des paramètres (cf. Figure 4.1).

4.1 Problématique

Lorsqu'une méthode d'estimation est développée, il est nécessaire d'en évaluer les performances pour identifier les conditions qui permettent d'obtenir des estimations fiables. Les publications qui présentent de nouvelles méthodes contiennent généralement une analyse de sensibilité à partir de données simulées et une application à des données réelles. L'étude à partir de données synthétiques est souvent limitée et requiert d'être approfondie. Par ailleurs, la validation des méthodes avec des données réelles n'a de sens que si les paramètres sont connus - i.e. mesurés selon une autre approche, fiable - pour pouvoir comparer avec les estimations. Du point de vue de l'utilisateur, il est important de connaître (i) la qualité des estimations selon les régions de l'espace des paramètres, (ii) la robustesse de la méthode lorsque les données ne vérifient pas exactement les hypothèses du modèle.

Dans leur article, Wilson and Rannala (2003) ont validé leur méthode avec des données simulées selon le modèle d'inférence implémenté dans BAYESASS. Plus précisément, ils ont considéré un scénario avec deux populations, des taux de migration symétriques et des loci biallélliques. Leur analyse de sensitivité leur a permis d'étudier l'effet du niveau de différentiation génétique, du taux de migration, du nombre de loci et d'individus échantillonnés sur la qualité des estimations. Par ailleurs, leur étude présente une application à des données réelles chez une plante, la Centaurée de la Clape *Centaurea corymbosa* (Freville et al. 2002) et chez le loup

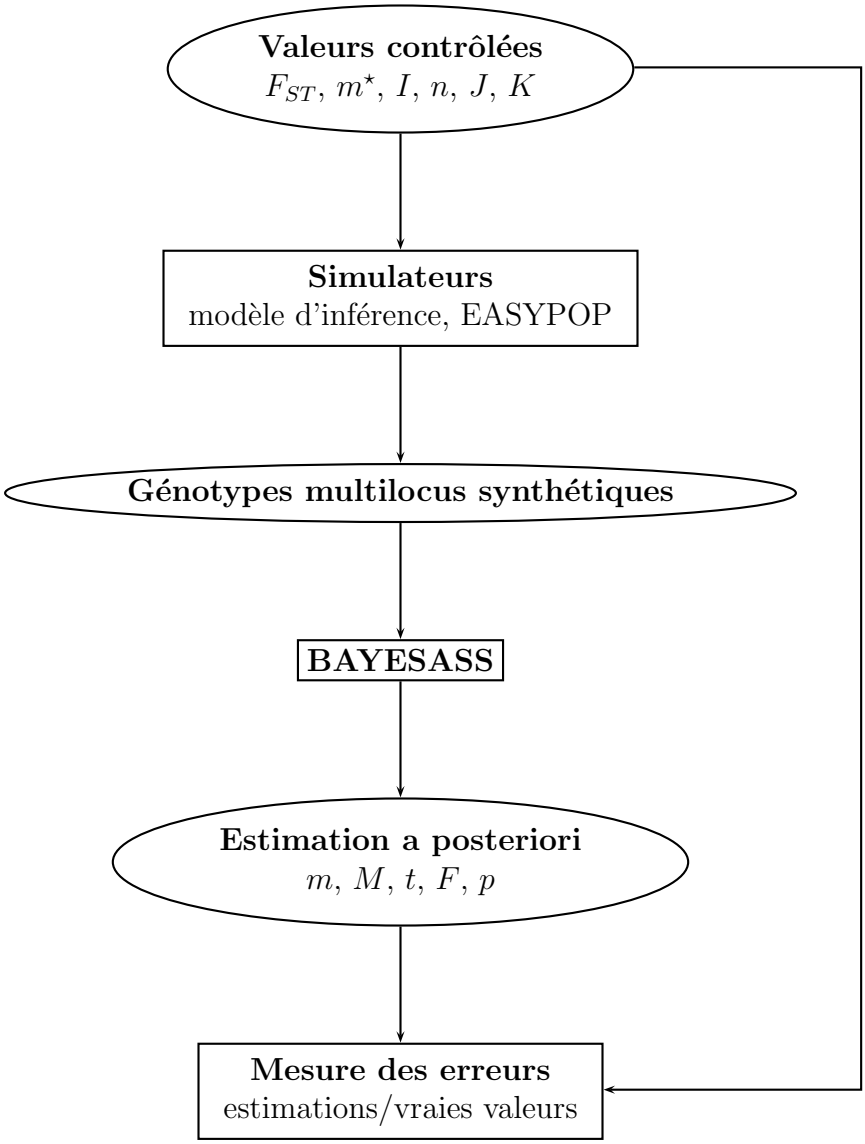


FIGURE 4.1 – Schéma du processus d'évaluation de BAYESASS.

gris *Canis lupus* (Carmichael et al. 2001).

Une évaluation plus poussée de BAYESASS permettrait une meilleure utilisation de la méthode pour obtenir de précieuses informations sur les flux géniques. Dans ce but, l'analyse de sensibilité doit explorer plus en profondeur l'espace des paramètres et envisager des scénarios qui s'écartent du modèle d'inférence de la méthode.

4.2 Modèle et méthodes

L'approche implémentée par BAYESASS permet d'estimer non seulement les taux de migration récents, mais aussi les fréquences alléliques, les coefficients de consanguinité et l'origine des individus. Tous ces paramètres sont intégrés dans un modèle bayésien hiérarchique dont les données sont les génotypes multilocus et les populations d'échantillonnage des individus. La méthode d'estimation repose sur l'algorithme de Metropolis-Hastings pour simuler la loi a posteriori des paramètres.

Le modèle de migration de BAYESASS est un modèle matriciel, les taux de migration entre sous-populations sont arbitraires et supposés constant dans le temps. Les deux principales hypothèses sur les flux de gènes sont que (i) l'échantillonnage a lieu après deux vagues de migration, (ii) les taux de migration sont faibles, négligeables au deuxième ordre. Ces quelques considérations impliquent que les taux d'immigration sont compris entre 0 et 1/3 (cf. Appendice A, Wilson and Rannala 2003).

L'évaluation de BAYESASS repose sur la simulation de jeux de données synthétiques selon deux scénarios de migration en îles (d'effectifs identiques ou variables). Le premier suit le modèle d'inférence, les génotypes multilocus sont générés à partir des fréquences allèliques obtenues à partir d'une loi de Dirichlet ou d'une loi Dirichlet-multinomiale qui dépend du niveau de différentiation génétique F_{ST}. Dans le deuxième scénario, les données sont simulées selon un modèle individu centré implémenté dans EASYPOP (Balloux 2001).

La maitrise des paramètres des deux méthodes de simulation utilisées permet de mesurer l'effet d'une variable sur les estimateurs a posteriori des taux de migration. En particulier, il est intéressant d'évaluer l'influence de paramètres biologiques (e.g. F_{ST}, taux de migration, nombre d'îles, effectifs des sous-populations) et de la quantité d'information à disposition (e.g. taille de l'échantillon, nombre de loci, polymorphisme)

4.3 Résultats

Dans la mesure où les hypothèses du modèle de BAYESASS sont vérifiées, la méthode de Wilson and Rannala (2003) est performante lorsque la différentiation génétique n'est pas trop faible ($F_{ST} \geq 0,05$). Dans le cas contraire, de bonnes estimations ont été obtenues lorsque les taux de migration sont faibles ($m = 0,01$) et la différentiation génétique élevée ($F_{ST} \geq 0,10$).

L'un des résultats de BAYESASS est que la distribution a posteriori obtenue à partir de données non-informatives ne correspond pas à la distribution a priori. En effet, lorsque les données contiennent peu d'information la loi stationnaire de la chaîne de Markov approche la loi a priori. Cependant, dans le cas où la différentiation génétique est faible ($F_{ST} < 0,05$), les données génétiques contiennent peu de renseignements sur la structure de la population. Or, dans ces conditions, BAYESASS produit des résultats qui indiquent une forte structuration spatiale au lieu de taux de migration égaux. Les raisons possibles de ce phénomènes sont expliquées dans la section qui suit.

4.4 Conclusions et perspectives

La méthode implémentée dans BAYESASS permet d'estimer les taux de migration récents ainsi que d'autres paramètres d'intérêt pour la biologie de la conservation, l'écologie ou la génétique. L'évaluation des méthodes permet aux utilisateurs de comprendre les limites des modèles et des techniques d'estimation pour mieux interpréter leurs résultats. En particulier, dans le cas de BAYESASS l'hypothèse de faibles flux de gènes semble très importante. Les futures méthodes d'estimation des taux de migration devront affaiblir ces hypothèses pour fournir des informations de plus en plus précises quelque soit l'espèce considérée.

Comme de nombreuses méthodes bayésiennes basées sur les techniques MCMC, BAYESASS présente des problèmes de convergence. Dans la littérature il existe différents critères - calculés à partir de la vraisemblance a posteriori - pour minimiser ce type de défauts mais leur utilisation ne permet pas de confimer la convergence de la chaîne de Markov. Il est donc recommandé aux utilisateurs de la méthode d'initialiser différemment plusieurs MCMC pour en comparer les sorties et analyser un jeu de données.

Le code source de BAYESASS est mis à disposition des utilisateurs et permet d'avoir accès à l'implémention de la méthode MCMC. Il est alors instructif de remarquer comment les paramètres sont susceptibles de changer à chaque itération de la chaîne de Markov :

- le coefficient de consanguinité dans une seule sous-population,
- les fréquences alléliques à un seul locus, dans une seule sous-populations,

– les taux de migration dans une seule sous-population,
– enfin un seul individu pourrait voir son assignation changer.

Ainsi, chaque mise à jour est peu couteûse mais ce type d'implémentation pourrait bien être la cause des problèmes de convergence de BAYESASS. En effet, le temps d'atteinte de la distribution stationnaire, notamment des fréquences alléliques (et par conséquent de tous les autres paramètres), dépend de la stratégie d'exploration de la méthode MCMC qui dans ce cas paraît lente. Il est également probable que l'initialisation de la chaîne de Markov à partir de distributions non-informatives amplifie les défauts de convergence. Bien que l'idée d'initialiser la chaîne de Markov selon des lois informatives issues des données soit contestable d'un point de vue strictement bayésien, elle constitue une réponse pragmatique aux défauts de convergence.

Dans les cas où la différentiation génétique est faible, il est intéressant de remarquer que tous les individus sont assignés dans une seule sous-population. Il semblerait que dès que les fréquences alléliques d'une sous-population atteignent la loi stationnaire, du fait des faible F_{ST}, de plus en plus d'individus sont assignés dans cette sous-population. Lorsque une majorité d'individus se retrouvent dans ce cas, tout se passe comme dans un modèle continent-îles avec $m = 1$. Il faut également noter que la sous-population en question change selon les valeurs initiales de la chaîne de Markov.

Malgrès ses défauts de convergence la méthode implémentée dans BAYESASS permet de produire des estimations fiables lorsque les conditions mentionnées plus haut sont vérifiées. Les utilisateurs de la méthode, sous réserve de suivre les conseils du manuel du logiciel, pourront obtenir des estimations des taux de migration de qualité.

Aux vus des niveaux de différentiation génétique des populations naturelles analysées par Wislon and Rannala (2003) et de l'évaluation de BAYESASS, les résultats obtenus sur *C. corymbosa* et *C. lupus* méritent d'être confirmés par d'autres approches. Par exemple, il faudrait analyser ces jeux de données avec des méthodes bayésiennes qui utilisent des modèles d'inférence différents de celui de BAYESASS (e.g. STRUCTURE, Pritchard et al. 2000 ; GENELAND, Guillot et al. 2005 ; BAPS, Corander and Marttinen 2006 ; TESS, François et al.). De manière générale combiner plusieurs approches apportent plus d'information et de poids aux estimations et permet une meilleure interprétation des résultats.

L'étude de l'implémentation de BAYESASS laisse penser que la méthode MCMC de Wilson and Rannala (2003) peut être améliorée sans changer de modèle bayésien. En particulier, tester d'autres modes de mise à jour et/ou modifier la façon d'initialiser la chaîne de Markov pourraient rêgler certains défauts de convergence. Par ailleurs, d'autres méthodes bayésiennes de la génétique des populations utilisent un modèle avec fréquences alléliques corrélées (e.g. STRUCTURE, Falush

et al. 2003 ; GENELAND, Guillot et al. 2005 ; GESTE, Foll and Gaggiotti 2006) introduit par Balding and Nichols (1997). Bien que controversée dans les cas où le nombre de sous-populations est inconnu, l'utilisation d'un tel modèle permet de placer un prior informatif sur les fréquences allélique lorsque la différentiation génétique est faible. L'apport de cette extension à BAYESASS pourrait être étudié pour évaluer si cela améliore les estimations.

La suite de chapitre

Chapter 5

Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates

Pierre Faubet, Robin Waples and Oscar E. Gaggiotti, Molecular Ecology 2007

Bayesian methods have become extremely popular in molecular ecology studies because they allow us to estimate demographic parameters of complex demographic scenarios using genetic data. Articles presenting new methods generally include sensitivity studies that evaluate their performance, but they tend to be limited and need to be followed by a more thorough evaluation. Here we evaluate the performance of a recent method, BAYESASS, which allows the estimation of recent migration rates among populations, as well as the inbreeding coefficient of each local population. We expand the simulation study of the original publication by considering multi-allelic markers and scenarios with varying number of populations. We also investigate the effect of varying migration rates and F_{ST} more thoroughly in order to identify the region of parameter space where the method is and is not able to provide accurate estimates of migration rate. Results indicate that if the demographic history of the species being studied fits the assumptions of the inference model, and if genetic differentiation is not too low ($F_{ST} \geq 0.05$), then the method can give fairly accurate estimates of migration rates even when they are fairly high (about 0.1). However, when the assumptions of the inference model are violated, accurate estimates are obtained only if migration rates are very low ($m = 0.01$) and genetic differentiation is high ($F_{ST} \geq 0.10$). Our results also show that using posterior assignment probabilities as an indication of how much confidence we can place on the assignments is problematical since the posterior probability of assignment can be very high even when the individual assignments are very inaccurate.

5.1 Introduction

The study of dispersal processes (colonization and migration) is central to the fields of population genetics, molecular genetics and conservation and management of wildlife. Direct estimates of migration parameters can be obtained using purely ecological approaches such as mark-release-recapture methods (MRR), but they have the inconvenience of being time consuming and impractical for the study of large and/or spatially extended metapopulations. Indirect methods based on population genetics models are an attractive alternative because they are easy to implement in these situations and only require a carefully planned sampling programme aimed at collecting tissue samples for DNA extraction and analysis. For many decades these estimates were obtained from F-statistics, but more recently this practice has come under criticism due to the simplistic assumptions (constancy in demographic parameters and genetic equilibrium conditions) made by this approach (e.g. Whitlock & Mccauley 1999). Recent progress in population genetics theory and statistics has led to the development of sophisticated methods that avoid many (and sometimes most) of these unrealistic assumptions, and there is a growing interest in applying them to address practical questions in conservation and evolution.

Methods aimed at estimating migration parameters can be grouped into two types of approaches: (i) coalescent or genealogical approaches that use the genealogical information contained in DNA sequences, and (ii) multilocus genotype approaches that use gametic disequilibrium information. It is important to realize that these two types of methods differ not only in the type of information they use but also in the nature of the parameters they estimate. Coalescent methods (and those based on summary statistics) estimate long-term evolutionary parameters, while multilocus genotype methods estimate short-term ecological parameters.

It is a standard practice to publish the statistical genetic method with a limited validation study that is usually followed by a much more detailed one. This has indeed been the case for MIGRATE (first published by Beerli & Felsenstein 2001 and later evaluated by Abdo *et al.* 2004), the most well-known coalescent method for estimating migration rates. Here we evaluate the performance of a more recent method, BAYESASS (Wilson & Rannala 2003), which is the multilocus genotype counterpart of MIGRATE. It is based on a Bayesian approach and can estimate rates of recent immigration among populations. It also estimates the posterior probability distribution of individual immigrant ancestries, population allele frequencies and population inbreeding coefficients.

One of the most enticing features of Wilson & Rannala's (2003) method is that it has the potential for estimating contemporary migration rates among populations. It can thus be extremely useful for guiding conservation plans requiring the identification of demographically independent subpopulations. There is a paucity

of studies that address the question of how small migration rates (m) should be to insure that subpopulations have independent dynamics (Waples & Gaggiotti 2006) but a study by Hastings (1993) suggests that two populations become demographically independent when m falls below about 0.10. The preliminary simulation study of Wilson & Rannala (2003) suggests that their method might be capable of accurately estimating migration rates of this order of magnitude, but a more thorough evaluation is required to confirm this possibility.

In their sensitivity study, Wilson & Rannala (2003) considered biallelic markers and a scenario with two populations and investigated the effect of varying migration rates (0.01, 0.05, 0.10, or 0.20) and F_{ST} (0.01, 0.10, or 0.25). They also studied the effect of varying sample sizes (20 or 100 individuals) and number of loci (5, 10 or 20). Here we expand this simulation study by considering multiallelic markers and scenarios with varying number of populations. We also investigate the effect of varying migration rates and F_{ST} more thoroughly in order to identify more precisely the region of parameter space where the method is and is not able to provide accurate estimates of migration rate. We studied the effect of deviations to the assumptions of BAYESASS by generating data using both the same approach as Wilson & Rannala (2003) and another method, EASYPOP, which simulates a different biological scenario.

5.2 Methods

BAYESASS implements a Bayesian approach using Markov chain Monte Carlo (MCMC) techniques. In the next two sections, we describe the probabilistic model implemented by BAYESASS and the simulation techniques we used to generate the synthetic data. We also provide details of the parameters used in the MCMC runs and the statistics used to evaluate the performance of the method.

5.2.1 BAYESASS

The inference model implemented by BAYESASS assumes linkage equilibrium but allows for deviations from Hardy-Weinberg equilibrium by estimating population-specific inbreeding coefficients. Migration rates among populations can be asymmetric but are constant over short periods of time (two generations). Additionally, it is assumed that migration rates are small (see Appendix A in Wilson & Rannala 2003). These two latter assumptions impose a constraint on the range of migration rates that can be considered by the method. More precisely, the total proportion of migrant individuals into a population per generation cannot exceed 1/3. Thus, nonmigrant proportions must be in the interval 2/3 to 1. The method also assumes that genetic drift and migration during the last few generations do

not change subpopulation allele frequencies.

The Bayesian formulation implemented by Wilson & Rannala's (2003) method is,

$$f(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{p}|\boldsymbol{X}, \boldsymbol{S}) \propto \Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{p}) \times \Pr(\boldsymbol{M}, \boldsymbol{t}|\boldsymbol{m}) f_m(\boldsymbol{m}) f_F(\boldsymbol{F}) f_p(\boldsymbol{p}) \tag{5.1}$$

where the parameters to be estimated are $\boldsymbol{m} = \{m_{ql}\}$, a matrix of migration rates between populations, $\boldsymbol{F} = \{F_l\}$ a vector of inbreeding coefficients, $\boldsymbol{M} = \{M_h\}$, a vector that contains the source of migrant ancestry of individuals in the sample, $\boldsymbol{t} = \{th\}$, a vector that gives the generation at which migrant ancestors of the sampled individuals arrived, and $\boldsymbol{p} = \{p_{lji}\}$, a matrix with the subpopulation allele frequencies. The estimation is based on the multilocus genotypes $\boldsymbol{X} = \{X_{hj}\}$ and population source $\boldsymbol{S} = \{S_h\}$ of individuals in the sample.

The prior densities $f_m(\boldsymbol{m})$, $f_F(\boldsymbol{F})$, $f_p(\boldsymbol{p})$ and $\Pr(\boldsymbol{M}, \boldsymbol{t}|\boldsymbol{m})$, and likelihood function $\Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{p})$, are given in Wilson & Rannala (2003). The inference model is represented by the directed acyclic graph (DAG) in Fig. 5.1.

By default, BAYESASS provides means and variances of the parameters being estimated. In our case, we modified the program code in order to obtain the raw MCMC output and used R to estimate the probability density function, mean and mode of each parameter. The mode was estimated as the value that corresponds to the maximum of the probability density function (pdf), i.e. the value with the highest posterior probability.

5.2.2 Generation of synthetic data for simulations

We use two different approaches to generate the synthetic data. In a first instance, we used the same approach as Wilson & Rannala (2003), in which the simulation model follows very closely the inference model. This allowed us to carry out a detailed sensitivity analysis of the method. In order to investigate how the method performs when the scenario considered deviates from the inference model, we also generated data using the software EASYPOP (Balloux 2001).

Simulations of the inference model. We simulated samples from subpopulations exchanging migrants according to the Wright island model at stationarity. We considered the general situation of a species with discrete generations inhabiting I islands of constant size and studied J marker loci with K_j alleles at any given locus j (i.e. the number of alleles can vary among loci). Each generation a fraction m of the individuals on each island is replaced by immigrants from a large mainland population with constant allele frequencies $\boldsymbol{q} = \{q_{ji}\}$, where q_{ji} is the frequency of allele i at locus j. Under these assumptions, the stationary distribution of allele frequencies in the islands, $\boldsymbol{p} = \{p_{lji}\}$, follows a Dirichlet distribution

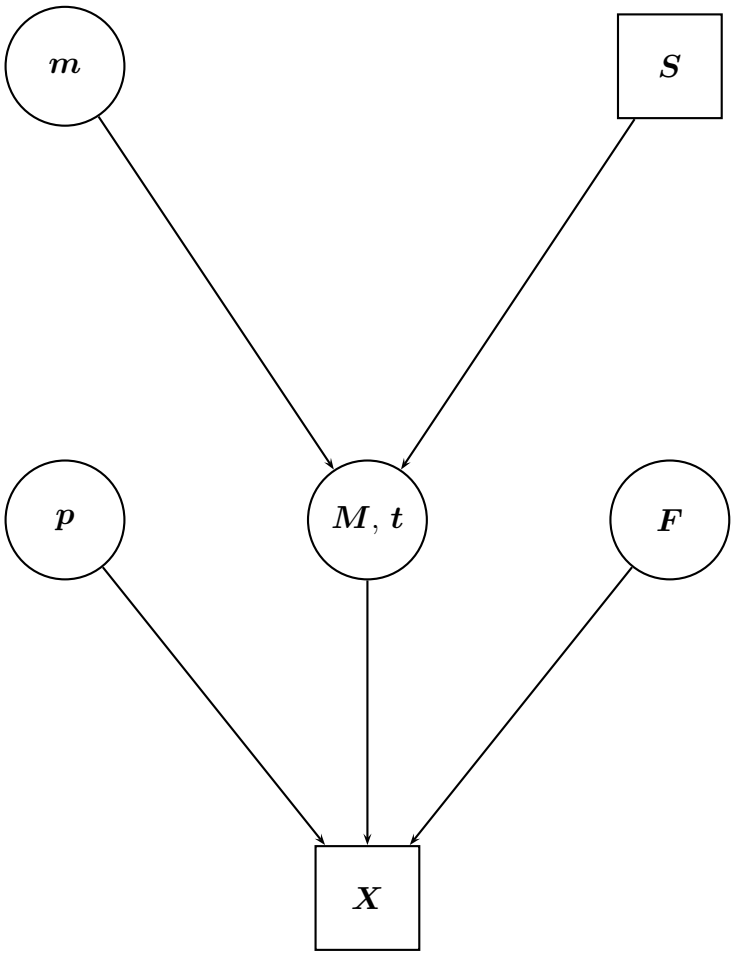


Figure 5.1: The Directed Acyclic Graph (DAG) for the model given in equation (5.1). Square nodes denote known quantities (data) and circles represent parameters to be estimated. Lines between nodes represent direct stochastic relationships within the model. The variables within each node correspond to the different model parameters discussed in the text.

with parameters $4Nm\boldsymbol{q}$, i.e.

$$f(p_{lj1}, \ldots, p_{ljK_j}) = \Gamma(4Nm) \prod_{i=1}^{K_j} \frac{p_{lji}^{4Nmq_{ji}-1}}{\Gamma(4Nmq_{ji})} \tag{5.2}$$

where p_{lji} is the frequency of allele i at locus j in population l, N is the (constant and equal) size of the subpopulations, and m is the proportion of migrants exchanged among populations.

In equation 5.2, $4Nm$ can be replaced by its expected value at stationarity, $4Nm \approx 1/F_{ST} - 1$, to obtain a pdf for generating allele frequency distributions at each locus and each local population with a fixed F_{ST} value. This approach does not allow the simulation of local populations that differ in size. Thus, to simulate this latter scenario we used the sampling formula for F_{ST} as described in Balding & Nichols (1997). The global allele frequencies, $\boldsymbol{q}$, used to generate the simulated data were those of the grey seal metapopulation studied by Gaggiotti *et al.* (2004), and only data sets with F_{ST}s that were within 10% of the targeted value were kept. We generated samples of n_q individuals from each simulated local population using the multinomial distribution of equation 2 in Wilson & Rannala (2003), which gives the probability of observing $\boldsymbol{M}$ and $\boldsymbol{t}$ given $\boldsymbol{m}$. To reduce the number of parameters to be considered in the simulations, we used symmetric and equal migration rates, i.e.

$$m_{lq} = m_{ql} = m^{\star} \quad \forall l \neq q \tag{5.3}$$

From equation (5.3) and the constraints on migrations rates imposed by the method (see above) we have

$$0 \leq m^{\star} \leq \frac{1}{3(I-1)} \tag{5.4}$$

The inference model assumes low migration rates and considers only possibilities involving at most a single migrant ancestor at some generation in the past. Thus, there are three types of individuals: nonmigrants, first generation migrants and second generation migrants (Wilson & Rannala 2003). The genotype of nonmigrants are generated by assigning alleles according to the Hardy-Weinberg proportions, conditional on the simulated allele frequency distributions of the population where the individual was sampled. Since the inference model assumes linkage equilibrium within each population, alleles are assigned independently at each locus. Genotypes of first generation migrants are generated according to Hardy-Weinberg proportions conditional on the allele frequencies in their population of origin. Second generation migrant genotypes are assigned by drawing an allele from each population.

Simulations using EASYPOP. We considered a finite island model with I subpopulations, each of constant size N and equal sex ratio. Each generation, random mating was simulated to produce a diploid genotype for J independent gene loci for each individual, which then had a probability m of migrating to another subpopulation. All loci had the same mutation dynamics, which occurred according to the k-allele model (KAM; each mutation equally likely to lead to any of k possible allelic states). We considered 10 allelic states and a mutation rate $\mu = 5 \times 10^{-4}$, values that are representative of highly polymorphic markers like microsatellites. Simulations were initiated with maximal genetic diversity (genotypes in initial generation randomly drawn from all possible allelic states). We ran each replicate for 5000 generations before collecting data to attain an approximate mutation-migration-drift equilibrium. In the final generation of each replicate, samples of n_q individuals were taken from each subpopulation for genetic analysis.

Accuracy and bias. We are particularly interested in the ability of BAYESASS to accurately estimate migration rates, but we also investigated the accuracy of the estimated inbreeding coefficients and the individual assignments. We used the posterior means and modes of the posterior distributions of m_{ql} and F_l as estimators of these parameters and evaluated accuracy using the relative mean square error (RMSE) for estimates of m_{ql} and the mean square error (MSE) for estimates of F_l. This was carried out in order to be able to compare the accuracy when varying migration rate. In the case of inbreeding coefficient, we limited ourselves to scenarios that assumed Hardy-Weinberg equilibrium ($F_l = 0$) so we use MSE instead of RMSE. We also calculated the relative bias for the estimators of migration rates and bias for estimates of inbreeding coefficient. In order to calculate these statistics, we simulated $N = 10$ independent data sets for each scenario and used the following equations:

$$RBias(\tilde{\boldsymbol{m}}) = \frac{1}{N}\frac{1}{I(I-1)}\sum_k\sum_{q\neq l}\frac{\tilde{m}_{lq}^k - m^\star}{m^\star} \tag{5.5}$$

$$RMSE(\tilde{\boldsymbol{m}}) = \frac{1}{N}\frac{1}{I(I-1)}\sum_k\sum_{q\neq l}\left(\frac{\tilde{m}_{lq}^k - m^\star}{m^\star}\right)^2 \tag{5.6}$$

$$Bias(\tilde{\boldsymbol{F}}) = \frac{1}{N}\frac{1}{I}\sum_k\sum_l \tilde{F}_l^k \tag{5.7}$$

$$MSE(\tilde{\boldsymbol{F}}) = \frac{1}{N}\frac{1}{I}\sum_k\sum_l (\tilde{F}_l^k)^2 \tag{5.8}$$

where $\tilde{m}_{ql}^k$ is the estimated migration rate from population l into population q obtained for the replicate data set k, and $\tilde{F}_l^k$ is the estimated inbreeding coefficients for population l obtained from data set k. Note that equations 5-8 give overall measures of bias and accuracy for the matrix of migration rates $\tilde{\boldsymbol{m}}$ and the vector of inbreeding coefficients $\tilde{\boldsymbol{F}}$, which are obtained by averaging across all the matrix/vector elements.

We obtained the 95% credible intervals (CI) for each element of the migration matrix and calculated its width. We also recorded the number of times that the true value fell within the CI. The results represent the average across all the elements of the migration matrix and replicates.

We evaluated the accuracy of migrant ancestry assignments using the proportion of individuals that were assigned to their correct (simulated) ancestral class and report the mean across all 10 replicates. We also use the maximum posterior probability with which these assignments were carried out. For each individual we recorded the population with the highest posterior assignment probability (irrespective of it being correct or false). These values were then averaged across all individuals in a data set and across all data sets. This was carried out only for data sets simulated under the inference model because in this case we knew how genotypes were drawn, which was not the case for data sets generated using EASYPOP.

Simulated scenarios. We chose a set of default values for the parameters of the simulation models and then studied the effect of varying only one of them at a time. For each simulation method and combination of parameters settings, we simulated 10 replicate data sets.

Table 5.1 presents the range of parameter values that we investigated with the simulations of the inference model. We looked at the influence of the level of genetic differentiation F_{ST}, number of individuals sampled per population n, number of loci J, number of alleles per locus K, number of populations I, and proportion of migrants $m = m^\star$.

Table 5.2 presents the parameter sets considered using EASYPOP. In this case, we investigated the effect of varying population sizes N, migration rates m, and numbers of populations I. The characteristics of the samples were kept constant: sample size of 50 individuals per population, 20 loci each with 10 allelic classes.

MCMC runs. We analysed the simulated data sets using MCMC runs of 21×10^6 iterations, to insure convergence. We discarded the first 10^6 iterations as burn-in and used a thinning interval of 2000 iterations. Instead of using the default values, we used delta values of 0.10 for all parameters because they resulted in acceptance rates that varied between 20 and 60%. We identified MCMC runs with

Table 5.1: Parameters for data generated with the inference model. The first column gives the parameter that was allowed to vary, and the range of values considered, the six that follow give the values assigned to the parameter that were fixed. The last column indicates the figure that show the results obtained for each scenario. F_{ST} = genetic differentiation, m = migration rate, I = number of populations, n = number of individuals sampled per population, J = number of loci, K = number of alleles per loci.

Parameter	True values considered	Fixed parameters						Fig.
		F_{ST}	$m^\star$	I	n	J	K	
F_{ST}	$\{0.01, 0.02, 0.05, 0.075, 0.1, 0.25\}$		0.05	3	100	10	11	5.2,5.4
$m^\star$	$\{0.0, 0.01, 0.02, 0.05, 0.1, 0.15\}$	0.1		3	100	10	11	
I	$\{2, 3, 5, 7\}$	0.1	0.05		100	10	11	
n	$\{20, 40, 60, 80, 100\}$	0.1	0.05	3		10	11	5.3,5.5
J	$\{5, 10, 15, 20\}$	0.1	0.05	3	100		11	
K	$\{2, 5, 8, 11\}$	0.1	0.05	3	100	10		

Table 5.2: Parameters for data generated with EASYPOP. We generated data for $J = 20$ loci with $K = 10$ possible allelic classes. A number of $n = 50$ individuals was sampled per population. I = number of populations, N = common population size, m = migration rate. The last column indicates figures where corresponding results are shown.

Parameter set	Input parameters				Fig.
	I	N	m	Nm	
m5	4	500	0.01	5	5.6(a), 5.7(a)
m1n2	4	200	0.01	2	5.6(a), 5.6(b)
m1n5	4	50	0.01	0.5	5.6(a)
m2n2	4	200	0.05	10	5.6(b)
m3n2	4	200	0.10	20	
m3n5	4	50	0.10	5	5.7(a)
n5	4	100	0.05	5	
2-25	2	500	0.05	25	5.7(b)
m25	4	500	0.05	25	
8-25	8	500	0.05	25	

convergence problems using two different approaches, depending on the method used to generate the data. In the case of data sets generated under the inference model, we considered as suspect any MCMC run that resulted in a very low proportion of individuals correctly assigned (less than 40%). In the case of data sets generated using EASYPOP, we focused on the quadratic error defined as $\sum_{q \neq l}((\tilde{m}_{lq}^k - m^\star)/m^\star)^2$ and considered as suspect any MCMC run that resulted in a quadratic error one order of magnitude larger than that of the best run (i.e. the one with the lowest error). We discarded MCMC runs with converge problems and repeated the analysis using different starting conditions until the proportion of individuals correctly assigned (for simulations under the inference model) or the migration rate quadratic error (for EASYPOP simulations) was the same order of magnitude as that of the best run. We also calculated the Bayesian deviance (see Appendix and Discussion) for all MCMC runs in order to establish if it could be used as a criterion to identify suspect runs when BAYESASS is applied to real data sets (see below). Low deviance values indicate a good fit of the data to the model (see Spiegelhalter *et al.* 2002) and therefore it may be possible to identify runs with convergence problems as those that lead to a high deviance.

5.3 Results

Here we present separately the results for the two types of data sets generated. We start by discussing convergence problems and then discuss the quality of the estimates using bias and RMSE. For each simulation method and parameter set (Tables 5.1 and 5.2), we plot relative bias (or bias) and RMSE (or MSE) of posterior means and modes. For data simulated under the inference model, we also plot proportion of individuals correctly assigned and assignment probability.

5.3.1 Simulations of the inference model

We detected convergence problems in 31 MCMC runs out of a total of 290. In these 31 cases, the MCMC chain got trapped in a region of high posterior probability and did not sample the whole parameter space, leading to estimates that deviated strongly from the true parameter values. We observed convergence problems more frequently for scenarios with low genetic differentiation ($F_{ST} = 0.01, 0.02$) or high migration rates ($m = 0.15$).

As explained in the Appendix, in the present case the Bayesian deviance can be decomposed into a term based on the likelihood of a genotype given a particular migration ancestry, D_{gen}, and a term based on the probability of a particular assignment given a migration rate, D_{assign}. For each replicate, we estimated both components and also the overall deviance. Interestingly, lack of convergence was

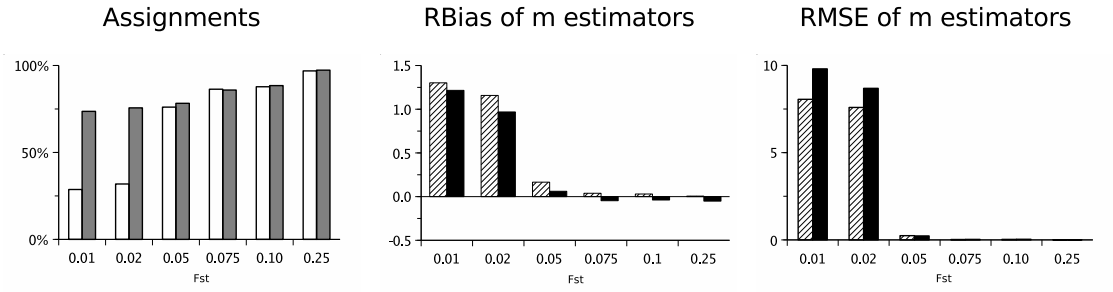
better identified using D_{assign} instead of the overall deviance. In all cases, D_{assign} of MCMC runs with convergence problems was much higher than that of 'good' runs (see Table 5.6, Supplementary material for an example), indicating that this statistic can be used for identifying suspect runs when the method is applied to real data sets.

The effect of genetic differentiation is very important; the accuracy of individual assignments and estimated migration rates increases with increasing F_{ST} values (Fig. 5.2(a)). Note that when genetic differentiation is low ($F_{ST} = 0.01, 0.02$), the individual assignments are very inaccurate but the maximum posterior probability with which individuals are wrongly assigned is very high. Thus, a high posterior assignment probability is not necessarily a good indication of how much confidence we can place on the assignments. As proportion of correct assignments increases, the bias of estimated migration rates decreases and their accuracy increases. In general, estimates of migration rates based on the mode are less biased than those based on the mean but their RMSE is larger, indicating that their variance is higher ($RMSE = RBias^2 + variance$).

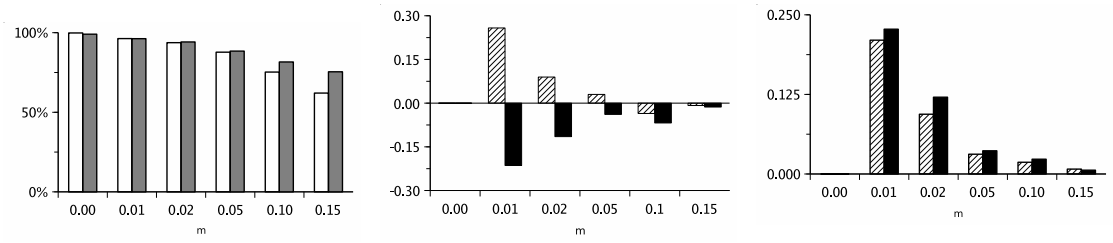
The effect of varying migration rates (Fig. 5.2(b)) is less pronounced than that of varying F_{ST}, probably due to the fact that $m^\star$ and F_{ST} are decoupled in these simulations. As migration rate increases, the proportion of correct assignments decreases but it is still above 60% for migration rates as high as 0.15 (when F_{ST} is fixed at 0.10). It is not possible to calculate the relative bias and RMSE when there is no migration ($m^\star = 0$), so for this particular case we calculated the bias and MSE (results not shown), which show that the mean produces overestimates while the mode has no bias at all. For low and intermediate migration rates, the mean gives overestimates while the mode gives underestimates; for large values both underestimate the true value. The bias and RMSE of both estimators decrease as migration rate increases. The observed change of sign in the bias of estimates based on the mean is due to the fact that the method sets an upper limit of 1/3 for the total proportion of migrants in a population. Thus, when the true migration rate is close to this upper limit, the parameter space becomes very asymmetric around the true value and the MCMC will visit more often smaller than larger values. This also has the effect of decreasing the RMSE because the MCMC will not be able to visit values that are much larger than the true value.

Increasing the number of populations decreases the accuracy of individual assignments and estimates of migration rates (Fig. 5.2(c)). With only two populations, bias is much larger for the mean than for the mode but as more populations are added, the bias of the latter increases rapidly while that of the mean decreases. The accuracy of both estimators of migration rates decreases rapidly as the number of population increases but more so for the mode than for the mean.

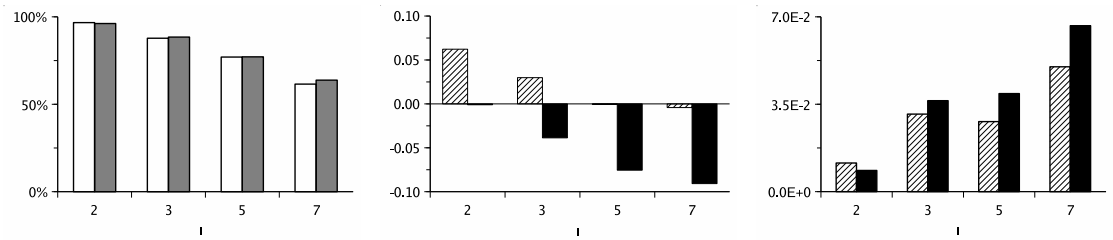
Another important aspect to investigate is the effect of size differences among



(a) Level of genetic differentiation F_{ST}



(b) Migration rate $m^\star$



(c) Number of populations I

□ = correct assignments
■ = maximum posterior probability of migrant ancestry assignments
▨ = posterior means
■ = posterior modes

Figure 5.2: Results for the data sets simulated under the inference model. Assignments, relative bias and RMSE of migration rate estimates when varying (a) level of genetic differentiation keeping $m^\star = 0.05$, (b) migration rate with F_{ST} fixed at 0.10 and (c) number of populations I with $m^\star = 0.05$ and $F_{ST} = 0.10$. Values of all other parameters are listed in Table 5.1.

local populations, since large differences are likely to increase the strength of genetic drift, and therefore have an effect on the accuracy of migration rate estimates. Table 5.3 compares the results for a scenario with equal local sizes (200 individuals) and another with two populations of size 50, one of size 200 and two others of size 500. When all local populations are equal in size, 77% of individuals are correctly assigned with a posterior probability of 0.77. However, when they differ in size, the proportion of individuals correctly assigned drops to 51% but the posterior probability remains high (0.75). The estimates of migration rate are also strongly affected and more so for the mean than for the mode. The relative bias of the mean for the scenario with unequal population sizes is one order of magnitude higher than that with equal sizes. Note that in the case of the mode, there is an underestimation of migration rates when all populations have equal sizes but an overestimation when they differ in size. The RMSE of both mean and mode is one order of magnitude larger when populations differ in size.

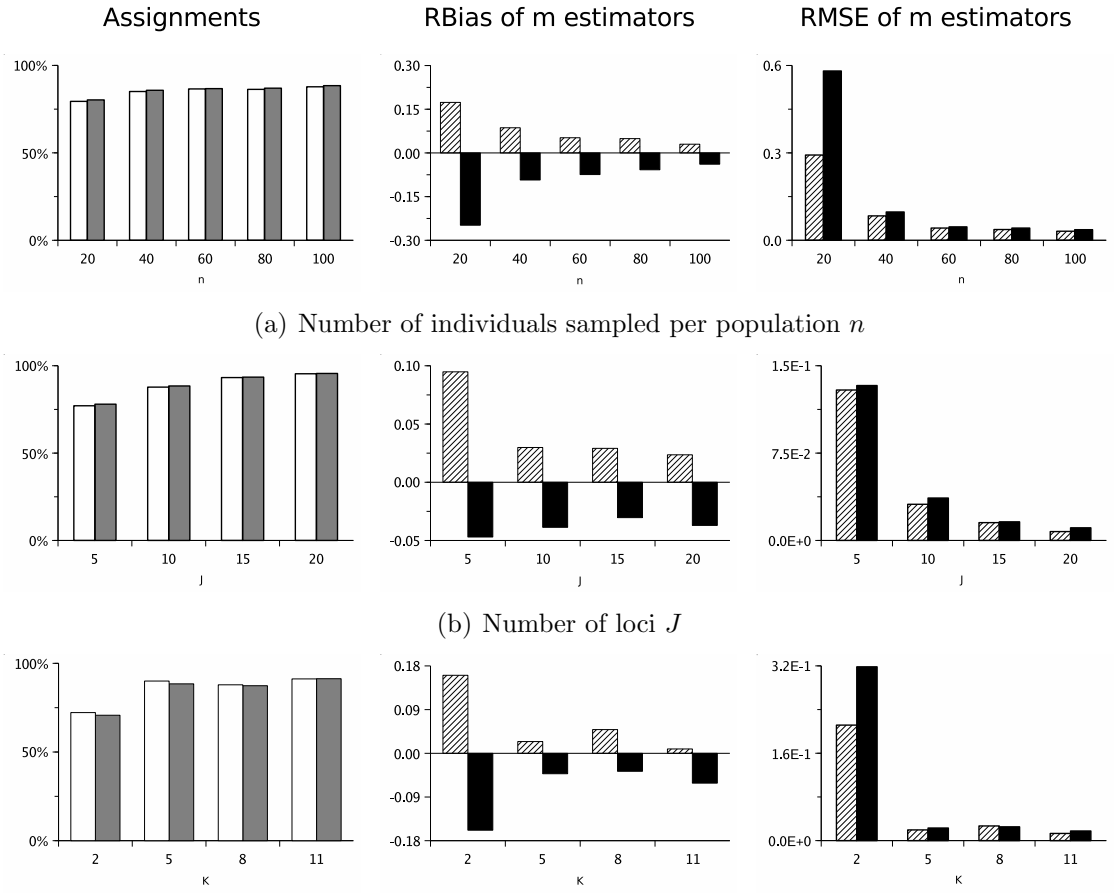
It is also important to investigate the effect of the amount of the data used for the estimation, which can be characterized in terms of sample sizes, number of loci scored and their degree of polymorphism (number of allelic classes). Sample size does not seem to have much of an effect on the accuracy of individual assignments, but this is not the case for estimates of migration rates (Fig. 5.3(a)). As sample size increases, the bias and RMSE of both the mode and the mean decrease. The mode always underestimates migration rates while the mean overestimates them, but the absolute value of the bias is more or less the same for both. The RMSE is much larger for the mode for sample sizes of 20 individuals, but for larger sample sizes it is the same as that of the mean. The quality of the estimates does not seem to improve a lot for sample sizes of 60 or more.

Increasing the number of loci increases the accuracy of the individual assignments and sharply decreases the bias of the mean but does not have much of an effect on the mode; on the other hand, the RMSE of both estimators decreases sharply initially but does not change much after 15 loci (Fig. 5.3(b)). Again, the mode underestimates the migration rates while the mean overestimates them. The accuracy of individual assignments is higher for multiallelic markers than for biallelic ones but not much is gained by using loci with more than five alleles (Fig. 5.3(c)). This is also true for the bias and RMSE of both the mode and the mean. As was the case before, the mode underestimates migration rates wile the mean overestimates them. It should be noted that these results correspond to a scenario with strong genetic differentiation ($F_{ST} = 0.1$); with lower F_{ST} values, accuracy is likely to continue to increase as the number of loci and their variability increases.

We also investigated the effect of varying the different model parameters on the width of credible intervals, CIs, of immigration rate estimates and on the proportion of times the true value fells within the CIs. As expected, increasing the

Table 5.3: Posterior estimates obtained when varying local population sizes in both simulation schemes with overlapping parameter spaces. We compare two scenarios: the first one with equal local sizes (200 individuals) and another with two populations of size 50, one of size 200 and two others of size 500. We report both relative bias and RMSE of mean and mode estimates and credible interval statistics.

		$RBias(m)$		$RMSE(m)$		95% CI width	Proportion of times true value falls within CI
Simulation scheme	Island sizes	mean	mode	mean	mode	Migration rate	Migration rate
Inference	Equal	$4.1E-01$	$-4.0E-01$	$1.2E+00$	$1.2E+00$	0.07	99%
model	Unequal	$1.2E+00$	$6.3E-01$	$1.2E+01$	$1.4E+01$	0.07	75%
EASYPOP	Equal	$8.5E-01$	$-1.2E-01$	$1.2E+01$	$1.1E+01$	0.07	66%
	Unequal	$1.2E+00$	$7.4E-01$	$1.8E+01$	$2.0E+01$	0.06	63%



(a) Number of individuals sampled per population n

(b) Number of loci J

(c) Number of alleles per locus K

- = correct assignments
- = maximum posterior probability of migrant ancestry assignments
- = posterior means
- = posterior modes

Figure 5.3: Results for the data sets simulated under the inference model. Assignments, relative bias and RMSE of migration rate estimates when varying (a) number of individuals sampled per population, (b) number of loci and (c) number alleles per locus. We fixed $F_{ST} = 0.10$ and $m^{\star} = 0.05$. Values of all other parameters are listed in Table 5.1.

information content of the data set (i.e. increasing F_{ST}, sample size, number of loci and/or number of alleles per locus) decreases the width of the CIs (Table 5.4). The proportion of times the true value is within the CIs is almost always 100%; only very low F_{ST}s (less than 0.05) can lead to much lower values. Increasing migration rates increases the width of the CIs but does not have an effect on the proportion of times they contain the true value (Table 5.4). The number of populations does not seem to have an effect on either measure (Table 5.4). Finally, size differences among local populations do not influence the width very much but it can greatly decrease the proportion of times the true value falls within the CI (Table 5.3).

Overall, these results indicate that if the assumptions of the inference model are not violated, the method can estimate migration rates fairly accurately when genetic differentiation is at least moderate ($F_{ST} \geq 0.05$) and samples are of good quality (40 individuals or more and 15 multiallelic markers). In general, it is preferable to use as estimator the posterior mean migration rate, which is more accurate than the mode of the migration rate (but see Discussion).

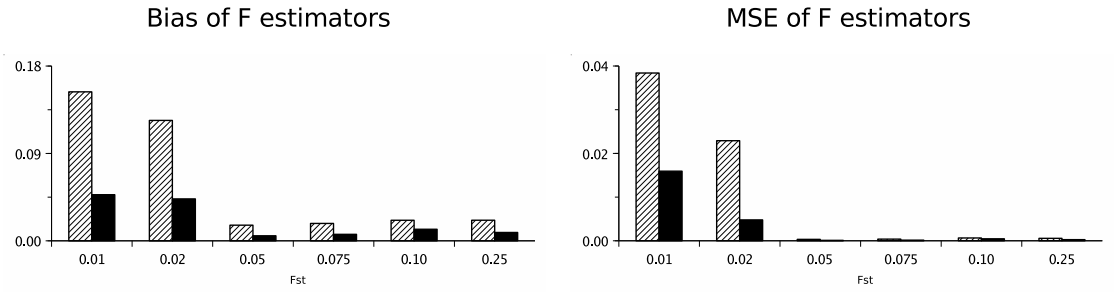
We investigated the quality of estimators of the inbreeding coefficient only for the case of scenarios with random mating within populations ($F_l = 0$) and therefore we report the results using the MSE instead of the RMSE (Figs 5.4 and 5.5). Contrary to what was observed for migration rates, the mode is a much better estimator than the mean because posterior distributions of F are very asymmetric. Also, both the mean and the mode overestimate F. As F_{ST} increases, the bias and MSE decrease, being fairly low for an F_{ST} of 0.05 or more (Fig. 5.4(a)). Increasing migration rates increases the bias and decreases the accuracy of the estimates of F (Fig. 5.4(b)). On the other hand, increasing the number of populations does not have much of an effect on the mode but does increase the bias and MSE of the mean (Fig. 5.4(c)). The effect of the quality of the samples on estimates of F is less important than for the estimates of migration rates (see Fig. 5.5). Increasing the sample size does improve the estimates based on the mean but does not have much effect on those based on the mode (Fig. 5.5(a)). A similar pattern is observed when increasing the number of loci (Fig. 5.5(b)). However, the effect of increasing the number of allelic classes is rather different, since the bias does not seem to depend on how polymorphic the markers are, while the MSE is much lower for multiallelic markers than for biallelic ones (Fig. 5.5(c)).

5.3.2 Simulations using EASYPOP.

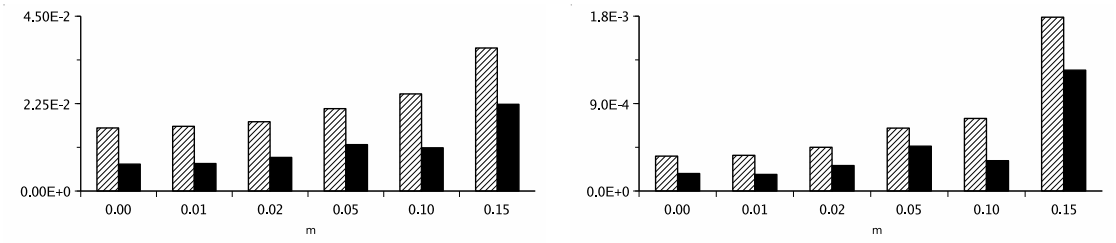
In the case of EASYPOP data sets, we observed convergence problems even for runs with the lowest quadratic error. We observed that the MCMC chain got trapped in regions that corresponded to the bounds of the prior distribution used for the migration rates. More precisely, the proportion of nonmigrants was either close to 2/3 or to 1; conversely, the proportion of immigrants from deme q into

Table 5.4: Credible intervals (CI) of migration rates for data simulated with the inference model. We report the width of the 95% CIs and the proportion of times the real value of the parameter falls within them when varying parameters.

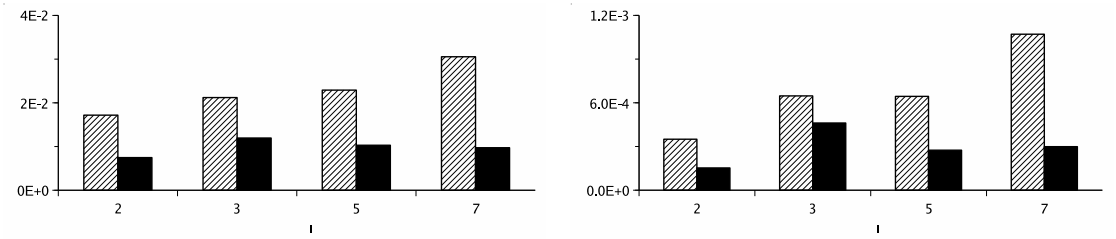
Parameter	Values	CI width	Proportion of times true value falls within CI	Fig.
F_{ST}	0.010	0.10	38%	5.2(a)
	0.020	0.07	27%	
	0.050	0.08	92%	
	0.075	0.06	100%	
	0.100	0.06	100%	
	0.250	0.05	100%	
$m^\star$	0.01	0.03	100%	5.2(b)
	0.01	0.04	100%	
	0.02	0.06	100%	
	0.10	0.07	100%	
	0.15	0.08	100%	
I	2	0.05	100%	5.2(c)
	3	0.06	100%	
	5	0.06	100%	
	7	0.06	100%	
n	20	0.13	100%	5.3(a)
	40	0.09	100%	
	60	0.07	100%	
	80	0.06	100%	
	100	0.06	100%	
J	5	0.08	100%	5.3(b)
	10	0.06	100%	
	15	0.05	100%	
	20	0.05	100%	
K	2	0.11	98%	5.3(c)
	5	0.06	100%	
	8	0.06	100%	
	11	0.05	100%	



(a) Level of genetic differentiation F_{ST}



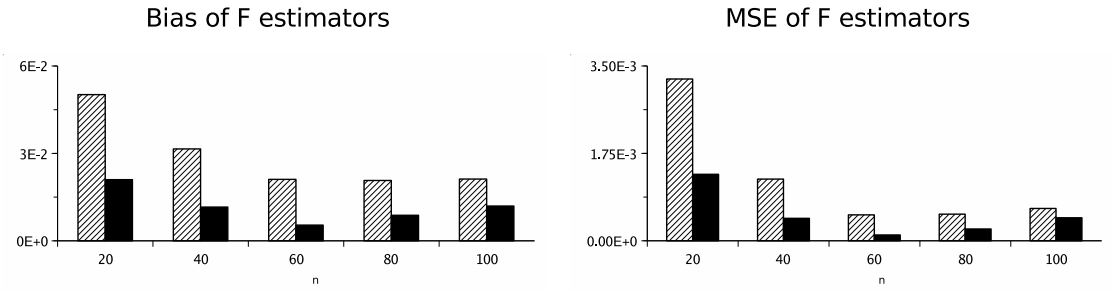
(b) Migration rate $m^\star$



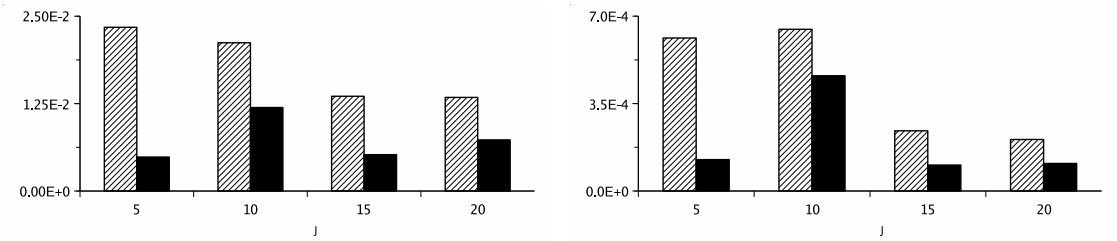
(c) Number of populations I

▨ = posterior means
■ = posterior modes

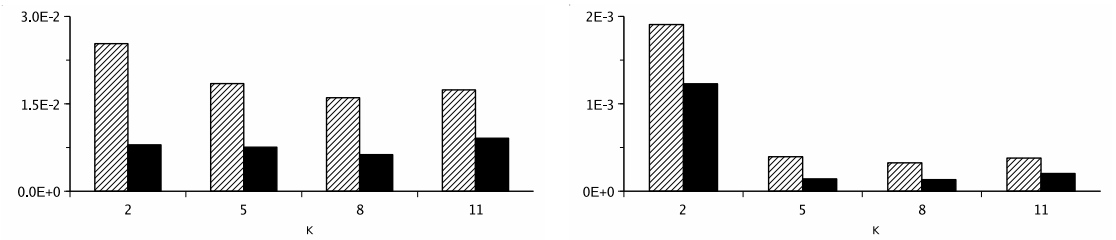
Figure 5.4: Results for the data sets simulated under the inference model. Bias and MSE of inbreeding coefficient estimates when varying (a) level of genetic differentiation keeping $m^\star = 0.05$, (b) migration rate with F_{ST} fixed at 0.10 and (c) number of populations with $m^\star = 0.05$ and $F_{ST} = 0.10$. Values of all other parameters are listed in Table 5.1.



(a) Number of individuals sampled per population n



(b) Number of loci J



(c) Number of alleles per locus K

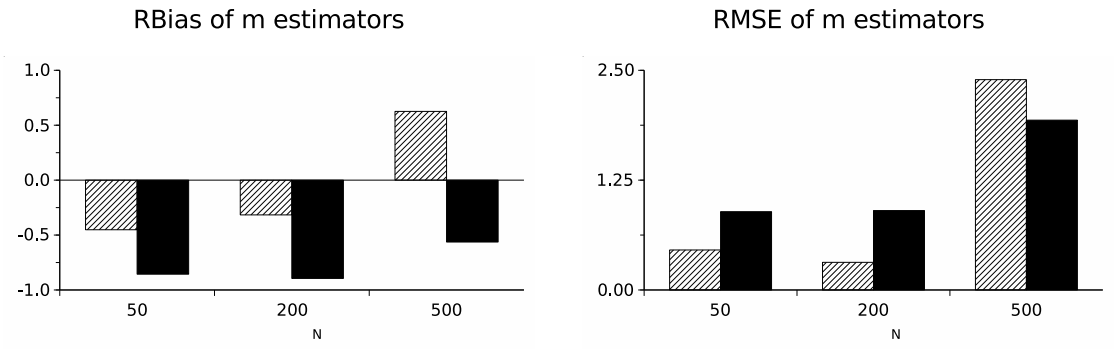
▨ = posterior means
■ = posterior modes

Figure 5.5: Results for the data sets simulated under the inference model. Bias and MSE of inbreeding coefficient estimates when varying (a) number of individuals sampled per population, (b) number of loci and (c) number alleles per locus. We fixed $F_{ST} = 0.10$ and $m^{\star} = 0.05$. Values of all other parameters are listed in Table 5.1.

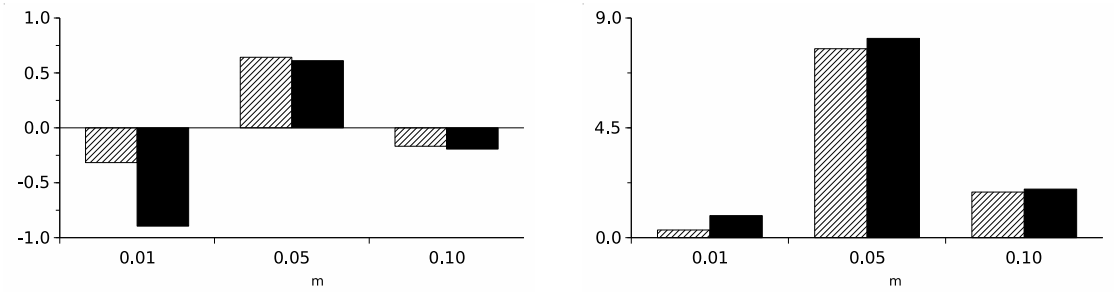
deme l was either very close to 0 or very close to 1/3 (see examples in Figure 5.8, Supplementary materials). The results we present in what follows correspond to MCMC runs that had a quadratic error of the same order of magnitude as the run with the lowest error of the corresponding scenario, but it should be noted that this does not guarantee convergence. Moreover, we found only one scenario (m1n2, see Table 5.2) for which the RMSE of data sets generated with EASYPOP is of the same order of magnitude as those observed for data sets generated under the inference model. This scenario corresponds to $N = 200$ and $m = 0.01$ in which case, the F_{ST} is high (0.11). The RMSE observed for all other scenarios are at least one order of magnitude larger than those obtained for data sets simulated under the inference model. It should be noted that even if there were convergence problems, the relationship between the quadratic error and the Bayesian deviance for the assignments, D_{assign}, was as expected, that is, runs with the lowest quadratic error had the lowest deviance (see Table 5.7, Supplementary material).

In the simulations of the inference model, we could fix F_{ST} and the migration rates separately because it is assumed that we start with subpopulations with a certain level of genetic differentiation, which then exchange migrants for two generations. In the case of EASYPOP, this is not possible since migrants are exchanged from the very beginning of the simulations and the degree of differentiation (at equilibrium) is determined by Nm, the effective number of migrants. Thus, increasing subpopulation sizes, N, while keeping migration rates fixed at $m = 0.01$, decreases genetic differentiation and this leads to an increase in bias and RMSE (Fig. 5.6(a)). Nm can also be increased by increasing m while keeping $N = 200$ constant. In this case, however, the results differ from those obtained when N increases and m is kept constant. As m increases, the relative bias and RMSE first increase and then decrease (Fig. 5.6(b)). If we keep Nm constant by increasing N while decreasing m, then relative bias increases while the RMSE first increase and then decrease (Fig. 5.7(a)). Thus, the quality of the estimates does not necessarily depend on F_{ST}. In fact, the explanation for these results (Figs 5.6(a) and 5.7(a)) is that, as mentioned before, convergence problems result in estimates of m_{ql} that tend to be either very close to 0 or very close to 1/3. Thus, the distance between the estimate and the true value is larger for $m^\star = 0.05$ than for $m^\star = 0.01, 0.10$. We also explored the effect of increasing the number of populations when the effective number of migrants per generation Nm equals 25 (Fig. 5.7(b)). As I increases, the bias and the RMSE of estimates based on both the mean and the mode decrease.

Finally, we explored the effect of unequal population sizes on migration rate estimates (Table 5.3). The relative bias of the mean increases with respect to that of the scenario with equal sizes but remains within the same order of magnitude. The bias of the mode goes from negative with equally sized populations to positive with unequal sizes. The RMSE of both mean and mode increases with unequal



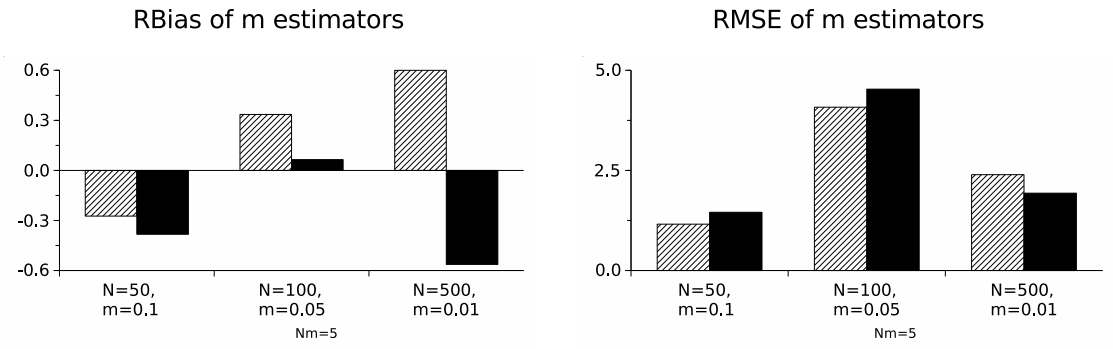
(a) Subpopulation sizes N



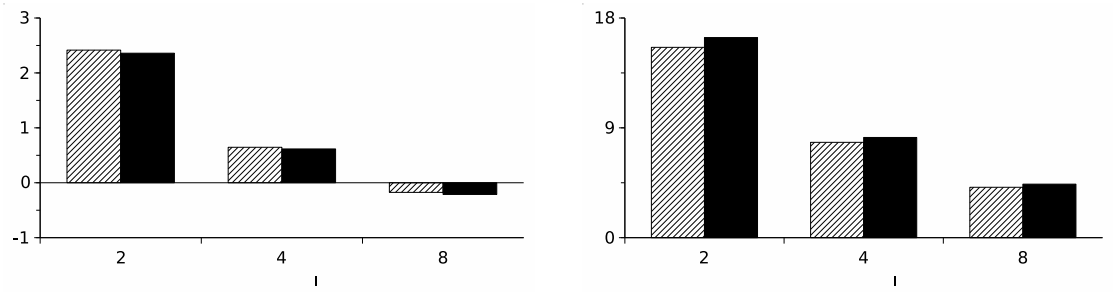
(b) Migration rate m

▨ = posterior means
■ = posterior modes

Figure 5.6: Results for the data sets generated using EASYPOP. Relative bias and RMSE of migration rate estimates when varying (a) subpopulation sizes while keeping migration rate constant ($m = 0.01$) and varying (b) migration rates while keeping subpopulation sizes ($N = 200$). Values of all other parameters are listed in Table 5.2.



(a) Effective number of migrants per generation Nm is fixed



(b) Number of populations I

▨ = posterior means
■ = posterior modes

Figure 5.7: Results for the data sets generated using EASYPOP. Relative bias and RMSE of migration rate estimates when varying (a) migration rate and subpopulation size while effective number of migrants per generation is fixed ($Nm = 5$) and varying (b) number of populations with $m = 0.05$ and $N = 500$. Values of all other parameters are listed in Table 5.2.

population sizes but remains within the same order of magnitude.

Varying m or N does not have much of an effect on the width of the CIs; on the other hand the proportion of times the true value falls within them is more sensitive to the migration rate than to the population size (Table 5.5). In particular, this latter measure is close to 0% for migration rates higher than 0.05. Keeping $Nm = 5$ while decreasing m and increasing N decreases the width of the CIs and increases sharply the proportion of times they contain the true value. Increasing the number of populations does not change much the width of the CIs while the proportion of time the true value falls within them is very low and drops to 0% when more than two populations are considered (Table 5.5). It should be noted here that Nm was kept constant at 25, which explains the low values observed for this measure. Finally, size differences among local populations do not change much either measures (Table 5.3).

EASYPOP does not allow the user to choose a fixed value for the inbreeding coefficient. Instead, it provides three choices for the mating system: random, polygyny, and monogyny. We chose random mating but it is clear that small populations will exhibit inbreeding even under random mating. Similarly, exchanging migrants can lead to a Wahlund effect increasing F. Thus, it is difficult to establish whether a positive bias in the estimates of this parameter is not in fact due to real inbreeding and Wahlund effects. For all these reasons, we do not present results for F for data sets generated with EASYPOP.

There is little agreement between the results obtained for the data sets simulated under the inference model and those generated using EASYPOP. For example, when varying the number of populations, the RMSE increased in the first case but decreased in the second. Moreover, as previously mentioned, there is only one scenario where the quality of the estimates obtained for EASYPOP data sets was similar to those observed for data sets simulated under the inference model. Even when parameters values were the same for both sets of simulations (Table 3), the quality of estimates were better for data simulated under the inference model than for data generated with EASYPOP. The deterioration due to unequal population sizes is more pronounced for data sets generated under the inference model than for those from EASYPOP. However, RMSE is still lower for the former than for the latter. These differences suggest that if the assumptions of the inference model are violated, the estimations of migration rate obtained should be interpreted with caution.

5.4 Discussion

The results indicate that if the demographic history of the species being studied fits the assumptions of the inference model, and if genetic differentiation is not too

Table 5.5: Credible interval (CI) of migration rates for data sets generated with EASYPOP. We report the width of the 95% CIs and the proportion of times the real value of the parameter falls within them.

Parameter	Values	CI width	Fig.
N	50	0.02	5.6(a)
	200	0.02	
	500	0.05	
m	0.01	0.02	5.6(b)
	0.05	0.02	
	0.10	0.02	
$Nm = 5$	$N = 50, m = 0.10$	0.08	5.7(a)
	$N = 100, m = 0.05$	0.07	
	$N = 500, m = 0.01$	0.05	
I	2	0.03	5.7(b)
	4	0.02	
	8	0.02	

low ($FST \geq 0.05$), then the method can give fairly accurate estimates of migration rates even when they are close to the threshold (about 0.1) that leads to correlated dynamics between populations. However, when the assumptions of the inference model are violated, accurate estimates are obtained only if migration rates are very low ($m = 0.01$) and genetic differentiation is high ($F_{ST} \geq 0.10$). Our results also show that using posterior assignment probabilities as an indication of how much confidence we can place on the assignments is problematical since the individual assignments can be very inaccurate but the maximum posterior probability with which individuals are (wrongly) assigned can still be very high, as illustrated by the results for scenarios with low genetic differentiation ($F_{ST} = 0.01, 0.02$; cf. Fig. 5.2(a)). This is a rather unexpected result since in principle, when genetic differentiation is low ($F_{ST} = 0.01, 0.02$), the actual conditional marginal likelihood function for different assignments is relatively flat. The true conditional posterior assignment probabilities should reflect this, by being pushed towards the uniform distribution among the local populations. However, BAYESASS results indicate high certainty in the assignments. A closer look at the MCMC output indicates that there is always a population with a very low immigration rate to which individuals from the other populations are assigned mainly as second-generation migrants. The other populations tend to have a proportion of nonmigrant individuals close to 2/3 (which corresponds to the lower bound of the prior distribution for m). Thus, the population to which an individual is assigned does not change much during an MCMC run, leading to a high posterior probability. It is important to note that as opposed to other methods such as structure (Pritchard *et al.* 2000), BAYESASS is not only carrying out assignments but is also estimating migration rates. Thus, the prior used for the migration rate can have an effect on the assignment of individuals. More precisely, although the prior for the vector of migration rates for any given population is uninformative, the marginal prior distribution for any given migration rate, m_{lq}, is not flat at all but L-shaped with a mode at zero (see Figure 5.9, Supplementary material). This type of prior might limit the mixing of the MCMC chain, forcing it to remain for very long periods of time at the same value of M (origin of migrant ancestor). Such a problem could be avoided by running extremely long MCMCs, or (probably more realistically) by improving the procedure with Metropolis-coupled MCMC (Geyer 1991).

We extended the simulation study of Wilson & Rannala (2003) by considering a larger number of populations and multiallelic markers. We also considered more values of migration rate and F_{ST} values. Our results confirm their suggestion that the use of multiallelic markers should increase accuracy of the estimates. However, as the number of populations increases, accuracy decreases. Within limits, increasing the quantity of information contained in the sample by increasing the number of loci and/or sample sizes also increases accuracy of estimates. Note,

however, that with up to three populations, not much is gained by using more than 15 loci and/or more than about 40 individuals.

The results of our simulations with EASYPOP suggest that the performance of the method is rather sensitive to deviations from the assumption of negligible change in allele frequencies due to migration and/or genetic drift over a few generations. This assumption is likely to be violated when migration rates are close to the 0.10 value considered as threshold for demographic independence. Thus, BAYESASS is unlikely to be useful for the identification of demographically independent units for borderline cases, which are the most interesting since it is very easy to identify demographic independence when migration rates are much smaller than 0.10.

The posterior mean of the migration rates seems to be a better estimator than the posterior mode because in general its RMSE is lower and its bias only a little bit higher than that of the mode. However, sometimes there is a need to be conservative. For example, we might prefer to err on the side of keeping two populations as separate management units rather than combining them; if this is the case, then it is better to use the mode. Additionally, when only two populations are involved (as is often the case in applications to management), the mode is always a better estimator of migration rates than the mean.

In general, although users of methods for the estimation of demographic parameters focus on point estimates rather than CIs, the latter can be a better way of evaluating the performance of Bayesian methods. In general, we expect that when data sets are highly informative, the width of the CIs will be narrow, while poor data sets will produce very wide CIs. In both cases, however, we expect that the proportion of times the true value falls within the CIs be very high. This is in general what we observe for the data sets simulated under the inference model; however, EASYPOP data sets give rather narrow CIs that in general do not contain the true value. This is another indication that we should be extremely cautious in the interpretation of results provided by BAYESASS when we suspect that the species being studied does not fit very well the assumptions of the method. One way of identifying unreliable results is to verify if the CIs are narrow and very close to one of the boundaries of the prior used for the migration rates. For example, one may obtain immigration rate estimates that are very close to 1/3 or 0, and correspondingly estimates of the proportion of nonmigrants that are close to either 2/3 or 1. If this is indeed the case, it is necessary to carry out many replicate analyses using very long MCMC runs (see below).

A practical problem associated with the use of MCMC is that of establishing whether or not the chain has converged. The basic principle implemented by the MCMC method is to construct an aperiodic and irreducible Markov chain whose stationary distribution (the 'target' distribution) is that given by the Bayesian

formulation (in our case eq. (5.1)). The estimation procedure consists of running the chain for 'sufficiently' long and treating the simulated values as a dependent sample from the target distribution (Brooks 1998). The underlying logic here is that the chain will visit more often regions of parameter space with a high posterior probability. In principle, the initial state of the chain (i.e. the initial values of the parameters we need to estimate) is arbitrary because we only start collecting data after the chain has reached equilibrium (i.e. converged). In practice, however, it is difficult to be sure that the chain has indeed converged. This is particularly the case with complex data sets and models, in which case the posterior probability is likely to be multimodal. The chain can then converge to one of the modes and remain in its vicinity for extremely long periods of time, giving the impression that it has converged. Running a second MCMC on the same data but with a different initial state can give very different results. Running longer chains is unlikely to solve this problem; for example, in our case we used runs of 21×10^6 iterations and still observed that many of them produced estimates that were very different from those obtained from runs that gave estimates very close to the true parameter values. In a simulation study such as ours it is easy to identify MCMC chains that did not converge because we know the true parameter values. However, in real applications this is not possible. One potential solution is to carry out multiple MCMC runs of the same data set and then compute a measure of model fit for each one of the runs, discarding those that provide a poor fit. One such measure of model fit is the Bayesian deviance (see References in Spiegelhalter *et al.* 2002); we explain how it is calculated in the Appendix.

In the present study, we analysed several replicates for each scenario and found that some posterior estimates departed strongly from the real values. Repeating the MCMC run on the same data set but with different initial conditions led to estimates that were much closer to the true values used as input for the simulations and to much lower statistical deviances. Thus, we suggest that in order to minimize convergence problems, it is advisable to carry out many MCMC runs, say 10, and select the one with the lowest deviance for obtaining the parameter estimates. Given that using extremely long MCMC runs does not seem to solve the convergence problem, we suggest using runs of 21×10^6, discarding the first 2×10^6 as burn-in. We have applied this strategy to one of the scenarios generated using EAYPOP (data set m1n2; see Table 5.2). Of the 10 replicate runs, three (6, 7 and 10) have a very high deviance for the assignment component, D_{assign} (Table S3, Supplementary material) and relative bias and RMSE at least one order of magnitude larger than the other runs (all of which have very similar low values). Table 6 presents the results taken from the best and from one of the worst runs (runs 1 and 6, respectively). In the chosen example, the true migration rate was 0.01 and the best run provides estimates that are almost identical to this value.

On the other hand, the estimates obtained from run 6 contain many migration rate estimates that are very different from the true value (m_{11}, m_{14}, m_{41} and m_{44}). In both cases, the CIs are very narrow regardless of whether the estimates are accurate or not. In this case, 10 replicates allowed discrimination between good and bad runs. However, if there are reasons to think that the species under study departs strongly from the assumptions of the inference model, then it would be appropriate to increase the number of replicate runs.

The most likely cause of the convergence problem we observed is the prior used for the migration rates, which sets bounds of 0 and 1/3 for the immigration rates and, equivalently, 2/3 and 1 for the proportion of nonimmigrants. Our simulation study shows that the chain gets trapped in regions of parameter space that correspond to these values. In this regard, it is interesting to note that in the example of the grey wolf provided by Wilson & Rannala (2003; Table 2), the estimated proportion of nonimmigrants into each population is close to either 1/3 or to 1. It is possible to avoid this convergence problem if the assumptions of the model (e.g. migration does not change the allele frequencies over the two generations considered) are not violated, in which case, following the advice provided above will suffice to insure convergence. However, if the assumptions are violated it is very difficult to avoid the biases introduced by the convergence problem.

Convergence problems have been reported for many recently developed Bayesian methods such as STRUCTURE, GENELAND (Guillot *et al.* 2005) and BAPS (Corander *et al.* 2004). In the case of STRUCTURE, Evanno *et al.* (2005) proposed a method based on running several MCMCs and calculating an ad-hoc statistic, Δk, based on the rate of change in the log probability of data between successive k values. The problem with this method is that there is always the potential of including in the calculation of Δk several chains that have not converged, leading to results that are unreliable. We observed this type of behaviour in a previous study (Waples & Gaggiotti 2006) and concluded that, for the simple finite island model that we considered, Evanno *et al.*'s (2005) method does not perform better than the original approach proposed by Pritchard *et al.* (2000). We think that it is better to use the same strategy used by Pritchard *et al.* (2000), namely run several chains for each value of k, say 20, and for each select the MCMC run that gives the smallest value of $-2\log \Pr(X|k)$. Using these chains one can then select the value of k that best fits the data set and base all estimations on the results of the best MCMC run. It should be noted that, as stated by Pritchard *et al.* (2000), $-2\log \Pr(X|k)$ is simply the mean of the Bayesian deviance penalized by a quarter of its variance.

In the case of GENELAND, Guillot *et al.* (2005) proposed a similar approach to that used by Pritchard *et al.* (2000) for STRUCTURE, but in this case they used the mode of the posterior distribution for the number of populations as the

criterion to choose the best MCMC runs. Finally, in the case of BAPS, Corander *et al.* (2004) proposed a similar strategy to that used by Pritchard *et al.* (2000) and Guillot *et al.* (2005) but using the posterior probability of the partition as the basis to select the best run.

Clearly, Bayesian methods such as the one we evaluate in this article are very powerful and offer an opportunity for answering difficult questions in ecology, population genetics, evolution and conservation biology, but we should be aware that their application is not as straightforward as that of the frequentist methods that have been used in past. Thus, users of these new methods should endeavour to follow very closely the recommendations provided by the software manuals and also seek the advice of colleagues competent in Bayesian methods. Furthermore, users should be aware that the models can have limited power to provide meaningful estimates under many realistic real-world scenarios, especially those that involve low levels of genetic differentiation.

5.5 Acknowledgements

We thank Bruce Rannala for helpfully discussing with us the results of this study. Three anonymous reviewers made very useful comments that greatly improved the manuscript. This work was supported by the Fond National de la Science (grant ACI-Impbio-2004-42-ADGP). P.F. holds a Ph.D. studentship from the Ministère de la Recherche.

5.6 Appendix: Bayesian deviance

In this section we outline the calculation of the Bayesian deviance, which we use to discriminate between MCMC runs that converged from those that did not. We base our discussion on the work of Spiegelhalter *et al.* (2002), who used the deviance statistic to define the DIC, a measure for choosing the model that provides the best fit among a group of alternative models. In our case we are not comparing models and therefore we simply use the Bayesian deviance, which has been proposed as a measure of model fit by a number of authors (see References in Spiegelhalter *et al.* 2002).

In Bayesian statistical modelling of data y we specify a prior distribution $f(\theta)$, $\theta \in \Theta$, and a likelihood $\Pr(y|\theta)$, which give rise to a marginal distribution

$$p(y) = \int_{\Theta} \Pr(y|\theta) f(\theta)\, d\theta \tag{5.9}$$

The Bayesian deviance is then defined as:

$$D(\theta) = -2 \log \Pr(y|\theta) + 2 \log g(y) \tag{5.10}$$

where $g(y)$ is some fully specified standardizing term which is function of the data alone. We can assume without loss of generality that $g(y) = 1$, so

$$D(\theta) = -2 \log \Pr(y|\theta) \tag{5.11}$$

We can thus estimate the expected deviance, $E_{\theta|y}[D(\theta)]$ from a MCMC run by taking the sample mean, $\overline{D(\theta)}$ of the simulated values of $D(\theta)$.

In order to calculate the deviance for a hierarchical model such as that implemented in BAYESASS, we need to define the parameter on which we want to focus. Hierarchical Bayesian models further parameterize the prior(s) with unknown 'hyper-parameters' ψ to obtain a full probability model

$$p(y, \theta, \psi) = p(y, \theta) \Pr(\theta|\psi) f(\psi) \tag{5.12}$$

Then, depending on the parameters in focus, we can specify the model in terms of the likelihood $\Pr(y|\theta)$ and prior $f(\theta) = \int_{\psi} Pr(\theta|\psi) f(\psi)\, d\psi$ or in terms of the likelihood $Pr(y|\psi) = \int_{\Theta} Pr(y|\theta) Pr(\theta|\psi)\, d\theta$ and prior $f(\psi)$. In our case, we are interested in using BAYESASS to estimate migration rates so we will focus on $\boldsymbol{m}$ and thus consider the likelihood, which is $\Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{p}) \Pr(\boldsymbol{M}, \boldsymbol{t}|\boldsymbol{m})$, and the prior $f_m(\boldsymbol{m})$. Thus, the deviance is composed of two terms, the first one, D_{gen}, concerns the likelihood of the genotypes and the second one, D_{assign}, the probabilities of assignments:

$$D(\boldsymbol{m}) = \underbrace{-2\overline{\log \Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{t}, \boldsymbol{F}, \boldsymbol{p})}}_{=D_{\text{gen}}} \underbrace{-2\overline{\log \Pr(\boldsymbol{M}, \boldsymbol{t}|\boldsymbol{m})}}_{=D_{\text{assign}}}$$

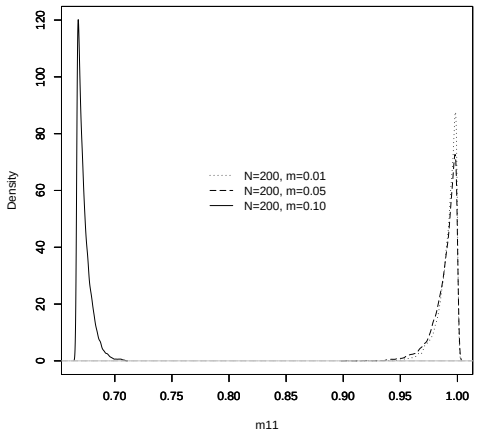
5.7 Supplementary material

Table 5.6: Results of the analyses of datasets simulated under the inference model. Simulated datasets consist of 10 repetitions with $I = 3$ populations with high level of genetic differentiation $F_{ST} = 0.10$ and low migration rate $m = 0.05$. We used $J = 10$ loci with $K = 11$ allele states and $n = 80$ individuals per population.

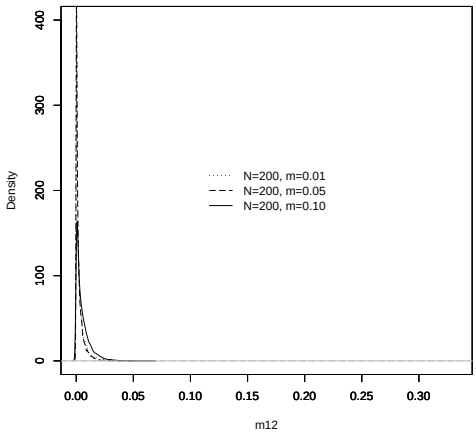
MCMC run	Proportion of individuals correctly assigned (%)	Relative bias of migration rate posterior mean	Deviance		
			Genotypes	Assignments	Total
01	87	estimator 3.83E-2	1 1215.98	51.13	1 1267.11
02	85	1.20E-2	1 0852.48	51.64	1 0904.12
03	**32**	**1.47E+2**	1 1025.84	**86.71**	1 1112.55
04	88	4.32E-2	1 1052.00	51.20	1 1103.20
05	87	1.33E-2	1 1099.80	53.07	1 1152.87
06	84	4.31E-2	1 1323.24	51.73	1 1374.97
07	88	6.84E-2	1 0910.00	51.60	1 0961.60
08	85	4.11E-2	1 0953.06	52.03	1 1005.09
09	88	3.71E-2	1 1058.38	51.60	1 1109.98
10	86	2.58E-2	1 0983.86	51.35	1 1035.21

Table 5.7: Results of the analyses of EASYPOP *m1n2* datasets. Simulated datasets consist of 10 repetitions with $I = 4$ populations of constant size $N = 200$ and low migration rate $m = 0.01$. We used $J = 20$ loci with $K = 10$ allele states and $n = 50$ individuals per population.

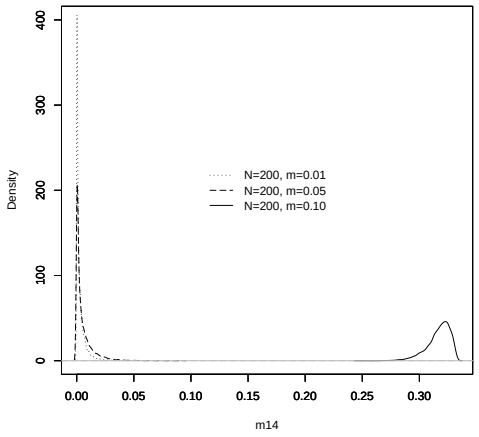
MCMC run	Relative bias of migration rate posterior mean		Deviance		
			Genotypes	Assignments	Total
01	estimator	-3.16E-1	1 5944.32	32.66	1 5976.98
02		-2.00E-1	1 3267.04	35.72	1 3302.76
03		**+4.61E+0**	1 4418.16	**120.58**	1 4538.74
04		**+4.62E+0**	1 4192.22	**120.92**	1 4313.14
05		-3.39E-1	1 4979.58	28.46	1 5008.04
06		-4.40E-1	1 3275.80	39.44	1 3315.24
07		-3.26E-1	1 4940.16	26.99	1 4967.15
08		-5.48E-1	1 4517.60	22.70	1 4540.30
09		-1.96E-1	1 5407.46	39.86	1 5447.32
10		-4.69E-1	1 3495.80	24.86	1 3520.66



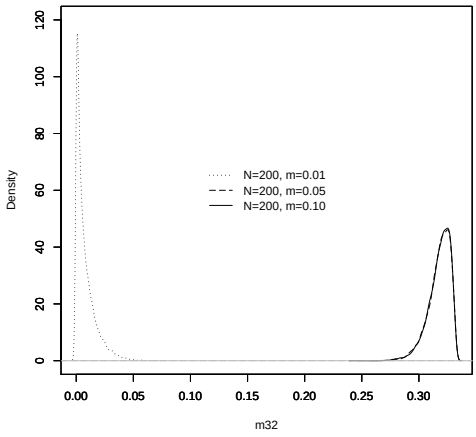
(a) Nonimmigrant proportion in population 1.



(b) Migration from population 2 into population 1.



(c) Migration from population 4 into population 1.



(d) Migrants from population 2 into population 3.

Figure 5.8: Results for data generated using EASYPOP. We represent posterior distributions of proportion of non-migrants (a) and immigration rates (b), (c) and (d) from typical output files corresponding to the three parameter sets presented in Fig. 5.6(b). Population sizes are fixed at $N = 200$ while varying m (0.01,0.05,0.10). The chains with migration rates equal to 0.05 or 0.10 are trapped in modes corresponding to the bounds of the prior for migration rates (0 and 1/3). This explain the pattern observed on figure 5.6(b) and 5.7(a) where RMSE is larger for a moderate value (0.05) of true migration rate than for extreme values (0.01, 0.10).

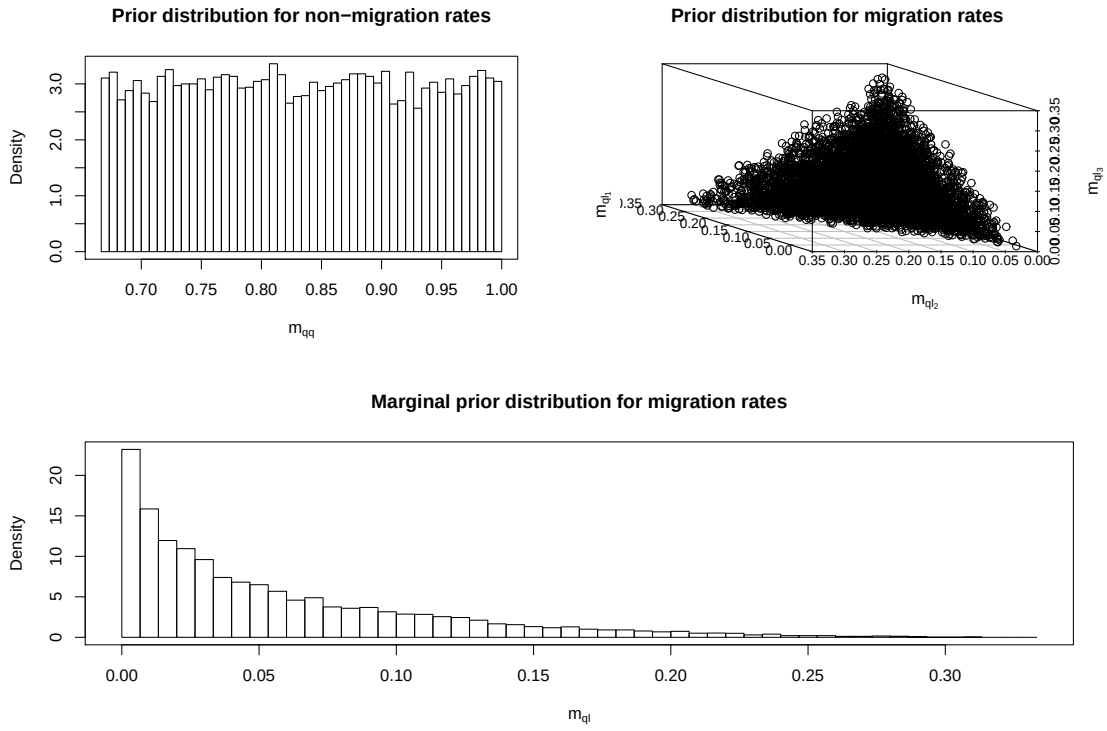


Figure 5.9: Prior distribution for migration rate matrix when considering $I = 4$ populations.

Chapitre 6

Article II

Ce chapitre présente l'article de Faubet and Gaggiotti (2008) dans lequel une nouvelle approche pour estimer les taux de migration récents est développée. La méthode est implémentée dans BIMr [1] et permet également d'identifier les facteurs environnementaux qui influencent les flux de gènes. L'approche est appliquée à des données simulées pour une étude de sensibilité et à des données réelles. Le logiciel qui implémente la méthode est disponible pour les trois principaux systèmes d'exploitation à l'adresse `http://www-leca.ujf-grenoble.fr/logiciels.htm`.

6.1 Problématique

Dans les populations naturelles, la migration entre sous-populations peut être influencée par des facteurs environnementaux. L'identification de ces variables et l'estimation de l'intensité avec laquelle elles agissent sont d'un grand intérêt pour la gestion des espèces. Différentes approches permettent de combiner données génétiques et coordonnées spatiales pour identifier les migrants (e.g. GENELAND, Guillot et al. 2005 ; TESS, François et al. 2006). Seulement, ces dernières ne sont pas conçues pour estimer directement les taux de migration et d'autres facteurs, biotiques ou abiotiques, que les distances géographiques peuvent agir sur la migration. Une autre approche consisterait à utiliser BAYESASS pour, dans un premier temps, estimer les taux de migration puis étudier leur éventuelle relation avec l'environnement.

Idéalement, la méthode estimera conjointement les taux de migration et le rôle des variables environnementales. Il s'agit donc d'étendre les possibilités de BAYESASS. Cependant, l'hypothèse des faibles flux migratoires de Wilson and Rannala (2003) (négligeables au second ordre) peut être jugée irréaliste pour certaines populations. De plus, la limite supérieure, de 1/3, imposée sur le taux global

1. Bayesian Inference of imMigration rates.

d'immigration dans chaque sous-population peut paraître insatisfaisante. Ainsi il semble nécessaire d'élaborer un nouveau modèle bayésien pour autoriser des taux de migration arbitraire.

L'approche retenue pour identifier les facteurs qui influencent la migration récente consiste à introduire les variables environnementales via la distribution a priori des taux de migration dans un modèle bayésien hiérarchique. Ce type de méthode s'est avérée particulièrement performante pourétudier l'influence de l'environnement sur la colonisation (Gaggiotti et al. 2002, 2004) et la différentiation génétique (Foll and Gaggiotti 2006).

6.2 Modèle et méthodes

Le modèle bayésien implémenté par BIMr est sensiblement différent de celui de BAYEASS. Les individus sont échantillonnés après la reproduction des migrants et les taux de migration peuvent varier librement entre 0 et 1. En plus des paramètres de BAYESASS, BIMr intègre un modèle avec fréquences alléliques corrélées qui permet d'estimer la différentiation génétique de chaque sous-population. La distribution a posteriori des paramètres du modèle bayésien est simulée à l'aide des méthodes MCMC.

La principale caractéristique de BIMr est que les paramètres du prior des taux de migration dépendent d'une régression linéaire qui tient compte de l'influence des variables environnememtales. Les coeffcients de la régression sont estimés pour mesurer l'intensité avec laquelle chaque facteur agit sur la migration. La considération de différentes combinaisons de variables environnementales permet d'envisager plusieurs modèles alternatifs dont les probabilités a posteriori sont estimées par une méthode RJMCMC.

Les performances de BIMr sont évaluées à partir de données simulées selon le modèle bayésien de la méthode. La génération de jeux de données synthétiques permet d'analyser la sensibilité selon la différentiation génétique, le nombre de populations, les taux de migration et la taille des échantillons. L'un des aspects importants est l'estimation des probabilités a posteriori des modèles et des taux de migration.

L'application de la méthode à des populations humaines (Cann et al. 2002) permettra de déterminer l'influence de l'altitude et des distances géographiques sur la migration au Pakistan. Le choix de ces populations est motivés par l'analyse de Rosenberg et al. (2002) qui a montré que ces populations appartennaient à deux groupes génétiques distincts.

6.3 Résultats

L'analyse de sensibilité à partir de données simulées montrent que BIMr produit des estimations de qualité dès lors que la différentiation génétique est strictement supérieure à 1%, le nombre de loci supérieur ou égal à 10 et la taille des échantillons supérieure ou égale à 50 dans chaque sous-population.

L'analyse des données génétiques des populations humaines du Pakistan fournit un exemple d'application lorsque la différentiation génétique est très faible. Les résultats obtenus indiquent que les différences d'altitudes influencent la migration et que les flux géniques sont importants dans les régions méridionales (zone de basse altitude) alors que les populations du nord semblent isolées (zone de haute altitude).

6.4 Conclusions et perspectives

La méthode développée permet de mesurer l'intensité des flux de gènes récents et de déterminer les facteurs environnementaux qui influencent la migration. Le modèle Bayésien de BIMr relaxe l'hypothèse des faibles flux de gènes de BAYESASS pour estimer des taux de migration récent arbitraires. De plus, l'utilisation de fréquences alléliques corrélées permet d'obtenir des informations sur les flux de gènes à plus long terme par le biais des niveaux de différentiation génétique des sous-populations.

Cependant, les modèles de BAYESASS et de BIMr ne font pas la distinction entre migration récente et admixture. En effet, la principale hypothèse de ces modèles est que les sous-populations sont d'abord à l'équilibre (avec un certain niveau de différentiation génétique) puis elles échangent des migrants pendant une ou deux générations. Les prochaines méthodes d'estimation des taux de migration devront modéliser le coefficient d'admixture plutôt que les taux de migration récents pour affaiblir encore les hypothèses des modèles.

L'implémentation des méthodes MCMC de BIMr est différente de celle de BAYESASS dans la mesure où, à chaque itération, sont mis à jour :

- les fréquences allélique de tous les loci, dans toutes les sous-populations,
- les coefficients de consanguinité de toutes les populations,
- les assignations de tous les individus,
- tous les taux de migration,
- l'influence de tous les facteurs environnementaux inclus dans le modèle,
- et enfin le modèle.

Cette implémentation est certes plus coûteuse que celle de BAYESASS mais elle corrige les défauts de convergence observés dans l'évaluation la méthode de Wilson and Rannala (2003). Notamment lorsque la différentiation génétique est faible (i.e.

données non-informatives), la distribution a posteriori correspond à la loi a priori.

Comme pour BAYESASS, BIMr requiert une évaluation plus approfondie pour permettre aux utilisateurs de la méthode de bien interpréter les résultats de leurs analyses. Il est tout aussi important de tester la robustesse de la méthode dans des scénarios qui ne respectent pas les hypothèses du modèle bayésien. Un autre aspect à étudier plus en détails avec des données simulées est l'apport de l'utilisation de fréquences alléliques corrélées pour prendre en compte les flux de gènes à long terme. Tous ces aspects pourraient être abordés avec EASYPOP qui permet de générer des populations selon différents modèles de migration à partir de simulations individus centrées.

L'un des aspects important, et qui est peu souvent étudié pour valider les algorithmes MCMC, est de déterminer la valeur prédictive du modèle et des estimations. L'idée serait de simuler des données selon le modèle de BIMr avec des paramètres a posteriori obtenus à partir de l'analyse d'un jeu de données réelles (e.g. les données humaines du Pakistan). Il s'agirait ensuite de comparer les analyses de ces jeux de données (réelles et simulées) avec une autre méthode (e.g. STRUCTURE).

L'un des principaux problèmes avec l'utilisation des méthodes telles que BAYESASS et BIMr est l'adéquation du modèle avec les données à analyser. Notamment, dans le cas des données humaines du Pakistan utilisées pour illustrer BIMr (ou du loup pour BAYESASS), les populations sont supposées discrêtes, concentrées en un point géographique. Cepandant, les travaux de Qamar *et al* (2002) et de Quintana-Murci *et al* (2004) ont montré une variation clinale de gènes du chromosome Y et d'ADN mitochondrial au Pakistan. Ceci pourrait constituer un biais que les modèles bayésiens de BIMr et de BAYESASS ne prennent pas en compte. Dans de prochaines études, l'utilisation de modèles spatialisés pourrait tenir compte des différences entre clines et clusters pour estimer les taux de migration.

L'analyse, avec BIMr, des jeux de données réelles utilisés pour illustrer BAYESASS permettrait de comparer les estimations produites par les deux méthodes. Notamment, il serait intéressant d'étudier les résultats de BIMr et ceux obtenus avec une régression directe des taux de migration de BAYESASS par les valeurs des facteurs environnementaux. Dans ce contexte, il serait possible de comparer différentes méthodes de sélection de modèles, de déterminer si l'utilisation des techniques RJMCMC est plus ou moins avantageuse qu'une approche basée sur le DIC/BIC ou les facteurs de Bayes.

La méthode BIMr a également été utilisée pour analyser des données chez l'escargot *Biomphalaria pfeifferi* (Charbonnel et al. 2002a, 2002b) et chez la grenouille *Rana latastei* (Ficetola et al. 2007). Dans les deux cas la différentiation génétique était élevée et indiquait une forte structuration spatiale. Les variables environnementales susceptibles de jouer un rôle étaient les distances géographiques,

l'appartenance à une même entité géographique (bassins versants, marres) ou la présence de possibles barrières à la migration. Les résultats obtenus (non présentés dans cette étude) montrent que, pour les deux espèces, les sous-populations sont totalement isolées avec de forts coefficients de consanguinité. Il faut noter que les analyses des mêmes jeux de données avec BAYESASS fournissaient des résultats similaires. L'absence de flux de gènes interdisait la mesure de l'influence des variables environnementales sur la migration et ne constituait pas une application statisfaisante de la méthode.

Chapter 7

A New Bayesian Method To Identify The Environmental Factors That Influence Recent Migration

Pierre Faubet and Oscar E. Gaggiotti, Genetics 2008

We present a new multilocus genotype method that makes inferences about recent immigration rates and identifies the environmental factors that are more likely to explain observed gene flow patterns. It also estimates population specific inbreeding coefficients, allele frequencies, local population F_{ST}s and performs individual assignments. We generate synthetic data sets in order to determine the region of the parameter space where our method is and is not able to provide accurate estimates. Our simulation study indicates that reliable results can be obtained when the global level of genetic differentiation (F_{ST}) is over 0.01 the number of loci is only 10 and sample sizes are of the order of 50 individuals per population. We illustrate our method by applying it to Pakistani human data considering altitude and geographic distance as explanatory factors. Our results suggest that altitude explains better the genetic data than geographic distance. Additionally, they show that southern low-altitude populations have higher migration rates than northern high-altitude ones.

The study of dispersal processes is an essential problem in ecology, population genetics, conservation, and management of wildlife. For this reason, the estimation of migration rates has been one of the most investigated problems in population biology. Migration parameters can be directly estimated using ecological approaches such as mark-release-recapture methods but they are not applicable to the study of large or extended metapopulations. In these cases, population genetics approaches provide a better alternative because the information contained in DNA can provide gene flow parameter estimates for different and complementary timescales. Methods based on coalescent theory provide long-term migration rates because they use the genealogical information contained in a sample of genes (e.g., MIGRATE, Beerli and Felsenstein 2001). On the other hand, methods based on multilocus genotypes (e.g., BAYESASS, Wilson and Rannala 2003) provide estimates of recent immigration rates by extracting the gametic disequilibrium signal generated by immigrant individuals or their descendants.

Besides simply estimating migration rates, it is very important to identify the biotic and/or abiotic factors that influence them. This can be done by first obtaining gene flow estimates and then searching for correlations between them and various environmental variables (e.g., Giordano *et al.* 2007). Such an approach requires the use of summary statistics that do not take advantage of all the information contained in genetic data. An alternative approach is to implement the joint analysis of genetic and nongenetic data. Several current methods that combine both genetic and geographic data can be used to detect recent migrants (e.g., GENELAND, Guillot et al. 2005; TESS, Francois et al. 2006) but they do not take into account other environmental factors.

In previous studies we presented methods that use genetic and environmental data to study colonization processes (Gaggiotti *et al.* 2002, 2004) and population genetic structure (Foll and Gaggiotti 2006). These approaches based on hierarchical Bayesian methods (e.g., Gelman *et al.* 1995) estimate the probability that a given environmental factor influences the parameters of interest (e.g., composition of colonizing groups or local population F_{ST}s) because they explicitly model the relationship between them and the relevant ecological factors. In this article we present a new multilocus genotype method for inferring recent immigration rates and identifying the environmental factors that best explain observed gene flow patterns. We use a hierarchical Bayesian approach that introduces nongenetic data through the prior distribution of the migration rates. Following Wilson and Rannala's (2003) approach we implement the estimation of inbreeding coefficients to allow for departures from Hardy-Weinberg equilibrium within local populations. Finally, the method infers the population ancestry of individuals by assigning their alleles to populations from which they originated. We carry out a simulation study to identify the region of parameter space where the method is and is not able to

provide accurate posterior estimates. We also illustrate our method with a real data example.

7.1 Data and model parameters

Inferring migration rates from genetic data: The method is based on a population genetics model that differs from that used by Wilson and Rannala (2003). More specifically, instead of assuming that sampling takes place right after migration, we consider that this is done after reproduction and before migration. Let us consider a metapopulation of a diploid species with nonoverlapping generations that is subdivided into I demes that can exchange migrants. Let $\boldsymbol{X} = (\boldsymbol{X}_{hl})$ be the observed multilocus genotypes of n individuals scored at L marker loci, where $\boldsymbol{X}_{hl}$ denotes the genotype of individual h at locus l. We assume that n_i individuals were sampled from population i and use the vector $\boldsymbol{S} = (S_h)$ to identify the population S_h where the individual h was sampled from.

Population allele frequencies are given by a matrix p composed of vectors pil that give the frequency of allele a at locus l for population i. Following Falush et al. (2003), we consider a model with correlated allele frequencies based on the approach introduced by Balding and Nichols (1995). Thus, we assume that before the last generation, the population was at migration-drift equilibrium so that allele frequencies in each population are determined by the global allele frequencies in the metapopulation as a whole, $\tilde{\boldsymbol{p}}_l = (\tilde{p}_{la})$, and the degree of genetic differentiation between each local population and the overall metapopulation, $\boldsymbol{\theta} = (\theta_i)$, where $\theta_i = 1/F_{ST}^i - 1$. Finally, to allow departures from Hardy-Weinberg equilibrium, we introduce population-specific inbreeding coefficients $\boldsymbol{F} = (F_i)$, where F_i is the inbreeding coefficient for population i. Thus, we consider two levels of inbreeding, one at the population level corresponding to F_{ST} and another one at the individual level, corresponding to F_{IS}.

Instead of focusing directly on individual migration rates, we consider the probability that genes in a deme originated in another one over the last generation. Thus, migration is described by a matrix $\boldsymbol{m} = (m_{ij})$, where m_{ij} is the probability that alleles in population i came from population j during the previous generation. The ancestral state of the individuals is described by a matrix $\boldsymbol{M} = (\boldsymbol{M}_h)$, where $\boldsymbol{M}_h = (i, j)$ is a two-element vector identifying the source demes (i and j) for the two alleles of individual h. All possible ancestry states are considered: both alleles come from the deme where the individual was sampled, or both come from another deme, or they come from two different ones. Thus, migration rates for individuals are obtained as

$$\tilde{m}_{ijk} = \begin{cases} m_{ij}^2 & \text{if } j = k \\ 2m_{ij}m_{ik} & \text{if } j \neq k \end{cases}, \quad j \leq k \tag{7.1}$$

where $\tilde{m}_{ijk}$ is the probability that individuals sampled from population i belong to the ancestry class (j, k). Note that our approach estimates migration rates only over the last generation. Moreover, as opposed to Wilson and Rannala (2003) migration rates vary freely in the interval $(0, 1)$ and do not have to be small.

The model parameters described above $(\boldsymbol{p}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{F}, \boldsymbol{M}, \boldsymbol{m})$ are estimated from the genetic data using a Bayesian approach and Markov chain Monte Carlo (MCMC) techniques.

Likelihood: The likelihood is the probability of the observed genotypes given model parameters and is constructed by defining the probability of observing the genotype of individual h at locus l in terms of the ancestry classes. We note these genotypes $\boldsymbol{X}_{hl} = (X_{hl1}, X_{hl2})$ where X_{hlc} is the allele observed at locus l in chromosome $c = 1, 2$ of individual h. Thus, individual h genotype likelihood at locus l is given by

$$\Pr(\boldsymbol{X}_{hl}|\boldsymbol{M}_h, \boldsymbol{F}, \boldsymbol{p}) = \begin{cases} \phi(\boldsymbol{X}_{hl}, i) & \text{if } \boldsymbol{M}_h = (i, i) \\ p_{ilX_{hl1}} p_{jlX_{hl2}} + \gamma p_{jlX_{hl1}} p_{ilX_{hl2}} & \text{if } \boldsymbol{M}_h = (i, j) \end{cases} \tag{7.2}$$

where

$$\phi(\boldsymbol{X}_{hl}, i) = \begin{cases} (1 - F_i) p^2_{ilX_{hl1}} + F_i p_{ilX_{hl1}} & \text{if } X_{hl1} = X_{hl2} \\ 2(1 - F_i) p_{ilX_{hl1}} p_{ilX_{hl2}} & \text{otherwise} \end{cases} \tag{7.3}$$

and

$$\gamma = \begin{cases} 0 & \text{if } X_{hl1} = X_{hl2} \\ 1 & \text{otherwise} \end{cases} \tag{7.4}$$

The first case considered in Equation (7.2) corresponds to the scenario where both alleles originated in the same source population, in which case we need to take into account possible deviations from Hardy-Weinberg equilibrium (see Equation (7.3)). The second case considers that the individual is the descendant of parents that come from two different source populations, in which case we need to take into account that there are two different ways of assigning the alleles to the parents.

If we assume that individuals were sampled at random and loci are unlinked, then the likelihood of the whole sample is obtained by multiplying across all loci and individuals,

$$\Pr(\boldsymbol{X}|\boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}) = \prod_{h=1}^{n} \prod_{l=1}^{L} \Pr(\boldsymbol{X}_{hl}|\boldsymbol{M}_h, \boldsymbol{F}, \boldsymbol{p}) \tag{7.5}$$

This likelihood can be used as the basis for inference using a Bayesian approach.

Combining genetic and environmental data: One can expect that migration patterns are influenced by environmental factors such as population densities, distances between local populations, etc. To identify which environmental factors have influenced gene flow we use Gaggiotti *et al.*'s (2004) approach. Let us suppose that our knowledge of the species under study leads us to think that R environmental factors $\boldsymbol{G} = (\boldsymbol{G}^{(r)})$ may influence the migration process. We can then introduce their effect through the prior distribution of gene migration rates. More specifically, we focus on the ancestry of immigrant alleles by conditioning on not being a resident allele

$$m_{ij}^{\star} = \frac{m_{ij}}{1 - m_{ii}} \tag{7.6}$$

and assume that the vector $\boldsymbol{m}_{\star i} = (m_{ij}^{\star})_{j \neq i}$ follows a Dirichlet distribution; i.e., $\boldsymbol{m}^{\star}{}_i | \boldsymbol{\psi}_i \backsim Dir(\boldsymbol{\psi}_i)$, where $\boldsymbol{\psi}_i = (\psi_{ij})_{j \neq i}$ are shape parameters for the Dirichlet distribution. Furthermore, we assume that each shape parameter ψ_{ij} follows a lognormal distribution; i.e., for each pair of distinct populations $i \neq j$

$$\log \psi_{ij} \backsim \mathcal{N}(\mu_{ji}, \sigma^2) \tag{7.7}$$

where the mean μ_{ij} is given by the generalized linear regression

$$\mu_{ij} = \alpha_0 + \sum_r \alpha_r G_{ij}^{(r)} + \sum_{r<s} \alpha_{rs} G_{ij}^{(r)} G_{ij}^{(s)} \tag{7.8}$$

where α_r denotes the effect of environmental factor r and α_{rs} denotes the effect of first-order interactions between factors r and s; these parameters are collected into a single vector $\boldsymbol{\alpha} = (\alpha_r, \alpha_{rs})$. The sign and the magnitude of the αs tell us about the direction and the strength of the environmental factors. Finally, σ^2 is the amount of variation that remains unexplained by the regression and G_{ij}^{r} is the observed value for factor r, which is hypothesized to influence migration between populations i and j. To reduce posterior correlation and to simplify prior elicitation and posterior interpretation process, explanatory factors are normalized before analysis so that they have zero mean and variance one.

By excluding different regression terms we can define different alternative models. We note, however, that as opposed to previous applications of this approach (cf. Gaggiotti *et al.* 2004; Foll and Gaggiotti 2006), the intercept α_0 is included in all models because it takes into account the effect of factors that act at a geographic scale larger than that of the metapopulation under study (see discussion for more details).

Other priors: We assume that there is no prior information on the shape of the other parameters and, therefore, adopt the vague priors that are given in the appendix. Note that in the particular case of the probability to observe

nonmigrant genes (i.e. m_{ii}), we adopt a uniform prior between 0 and 1 because, although some environmental factors may influence whether or not an individual decides to emigrate, our method is aimed at estimating immigration rates and, therefore, cannot take into account this possibility.

Posterior distribution: The model is now expressed in terms of parameters $\boldsymbol{\Theta} = (\boldsymbol{p}, \tilde{\boldsymbol{p}}, \boldsymbol{\theta}, \boldsymbol{F}, \boldsymbol{M}, \boldsymbol{m}, \boldsymbol{\psi}, \boldsymbol{\alpha}, \sigma^2)$ and the corresponding posterior distribution is given by Bayes' rule:

$$f(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{G}) \propto \underbrace{\Pr(\boldsymbol{X}|\boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}) \Pr(\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{m}) f(\boldsymbol{F}) f(\boldsymbol{p}|\tilde{\boldsymbol{p}}, \boldsymbol{\theta}) f(\tilde{\boldsymbol{p}}) f(\boldsymbol{\theta})}_{\text{Genetic data}} \times \underbrace{f(\boldsymbol{m}|\boldsymbol{\psi}) f(\boldsymbol{\psi}|\boldsymbol{\alpha}, \sigma^2, \boldsymbol{G}) f(\boldsymbol{\alpha}) f(\sigma^2)}_{\text{Environmental data}} \quad (7.9)$$

The full model is represented by the directed acyclic graph (DAG) in Figure . The posterior distributions of parameters given in Equation (7.9) are estimated using MCMC methods that are described in the supplemental information.

Posterior model probabilities: Besides estimating migration rates our method is aimed at identifying the environmental factors influencing gene flow. As we mentioned before, several alternative models can be obtained from the full regression 7.8 by canceling elements of the vector $\boldsymbol{\alpha}$. Note that models that include first-order interactions between factors r and s are allowed only if both factors are included. Thus for model $\mathcal{M}$, the corresponding posterior distribution is given by

$$f(\boldsymbol{\Theta}_{\mathcal{M}}|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{G}) \propto \Pr(\boldsymbol{X}|\boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}) \Pr(\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{m}) f(\boldsymbol{F}) f(\boldsymbol{p}|\tilde{\boldsymbol{p}}, \boldsymbol{\theta}) f(\tilde{\boldsymbol{p}}) f(\boldsymbol{\theta}) \times f(\boldsymbol{m}|\boldsymbol{\psi}) f(\boldsymbol{\psi}|\boldsymbol{\alpha}_{\mathcal{M}}, \sigma^2, \boldsymbol{G}) f(\boldsymbol{\alpha}_{\mathcal{M}}) f(\sigma^2) \Pr(\mathcal{M}) \quad (7.10)$$

where $\boldsymbol{\Theta}_{\mathcal{M}}$ is the parameter vector under model $\mathcal{M}$, $\boldsymbol{\alpha}_{\mathcal{M}}$ is the corresponding regression vector, and $\Pr(\mathcal{M})$ denotes prior model probability. Posterior model probabilities are estimated using the reversible-jump (RJ)MCMC approach (Green 1995, detailed in supplemental information). Here we note only that one of the problems faced when estimating posterior model probabilities is that the prior for σ_α^2 can have a large effect on the estimates. When very vague priors are used, more posterior weight is placed on the model with the fewest parameters. This is the well-known Jeffreys-Lindley paradox (Robert 1994). This problem was avoided by first running an MCMC of the full model with vague priors and then using the posterior estimates of the αs as informative priors for a new MCMC run.

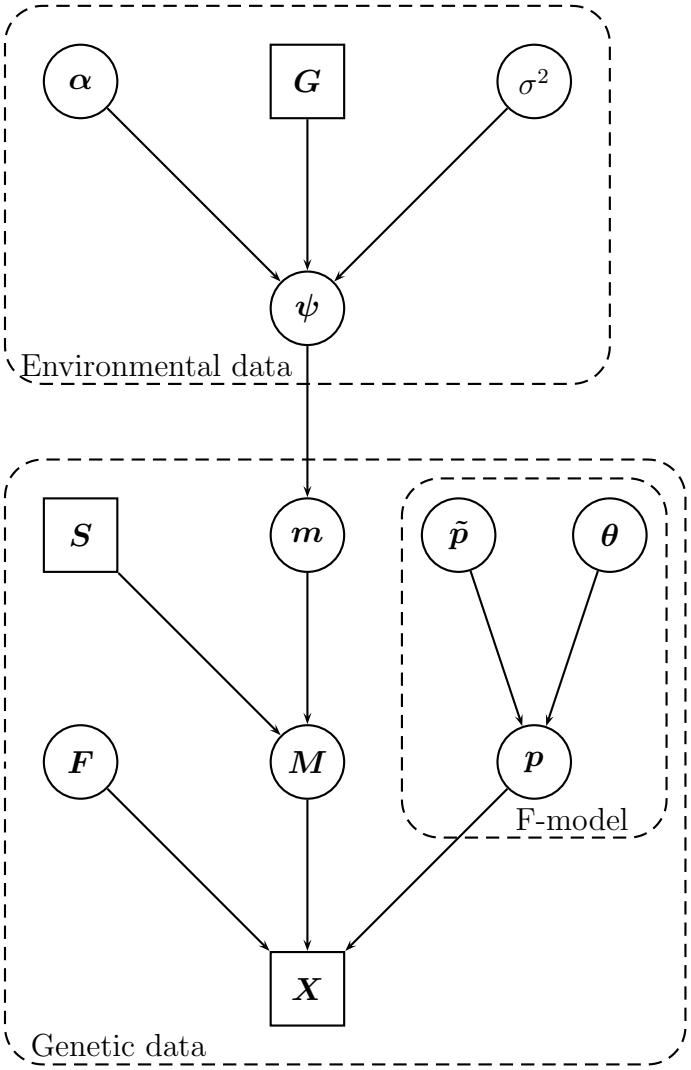


Figure 7.1: The DAG for the model given in Equation (7.9). Square nodes denote known quantities (i.e. data) and circles represent model parameters to be estimated. Arrows between nodes represent direct stochastic relationships within the model. The variables within each node correspond to model parameters discussed in the text.

7.2 Simulation study

We evaluated the sensitivity of our method by generating synthetic data under a particular scenario in which gene flow is influenced by two factors. We considered various levels of genetic differentiation and migration rates. We are interested in the ability of our algorithm to find the correct scenario and to provide accurate posterior estimates for migration rates (with corresponding fairly narrow highest posterior density intervals, HPDI).

Generation of synthetic data: We simulated data following the inference model presented in Figure 7.1. We initially considered a scenario with $I = 4$ populations, each with local population sizes of $N_i = 5000$ individuals. The sample size per population was $n_i = 100$ and we assumed that each sampled individual was scored for $L = 10$ polymorphic loci, each with $K = 10$ alleles.

Generating migration rates from environmental factors: We consider two environmental factors that could be, for example, geographic distance (G_1) and population density (G_2). The pairwise geographic distances were generated using a standard normal distribution. The pairwise differences in population density were generated by first filling the top triangular matrix with values drawn from a standard normal and then filling the bottom triangular matrix with the opposite values. This procedure is equivalent to standardizing the observed pairwise differences in environmental factors before analysis.

To generate the migration matrix, we first chose the values for the diagonal elements (proportion of non-migrant genes) and then we calculated the values for the nondiagonal elements (immigration rates) using the following procedure. We set $\alpha_1 = 0.9$, $\alpha_2 1.1$, and $\alpha_{12} = 0$ (i.e. no interaction effect) and calculated μ_{ij}s using Equation (7.8). Assuming no deviation from the linear regression (i.e. $\sigma^2 = 0$), we set $\psi_{ij} = e^{\mu_{ij}}$. Finally we computed the means $E[m^{\star}_{ij}|\boldsymbol{\psi}_i] = \psi_{ij}/\sum_{j\neq i}\psi_{ij}$ of the Dirichlet distribution used for the migration rates (Equation 7.13) and rescaled them so that they added up to $1 - m_{ii}$.

Genetic data: To generate multilocus genotypes with a given level of genetic differentiation, we need to generate parametric allele frequencies. This task was performed using Balding and Nichols' (1997) sampling formula for F_{ST}. According to this formula, genes are sampled one by one during an iterative process. The probability that the next gene sampled is a after having sampled n genes of which na correspond to allele a is given by

$$p_a(n_a, n) = \frac{n_a F_{ST} + (1 - F_{ST})p_a}{1 + (n-1)F_{ST}} \tag{7.11}$$

where p_a is the global frequency of allele a in the metapopulation.

To generate the parametric allele-frequency distribution of each locus in any given local population for a given F_{ST} value, we first sample an allele at random from the metapopulation allele-frequency distribution and then use Equation (7.11) to calculate the probability distribution for the type of the allele that will be sampled next. Using this distribution we obtain the next allele. This process is repeated iteratively until we obtain the $2N_i$ alleles present in local population i. We used uniform allele frequencies for the metapopulation.

Large departures from the target F_{ST} value were avoided by using the following iterative process. We generated the local population's allele-frequency distributions and calculated the global and pairwise F_{ST}s. If one or more of these values were not within 10% of the target value we discarded the allele frequencies and generated new ones. If the new allele frequencies satisfied the requirement, we generated the genotypes; otherwise we continued this iterative procedure until the constraint was satisfied. This procedure was used to control for the effect of genetic differentiation on the performance of the method.

Using the gene migration rates calculated above, we obtained the proportion of migrant individuals in each population using Equation (7.1). Multilocus genotypes for each local population were generated assuming Hardy-Weinberg equilibrium. Genotypes of nonmigrant individuals were obtained by drawing two alleles from the local population allele-frequency distribution. For the migrant individuals with both parents coming from the same local population, the genotype was obtained by drawing two alleles at random from the parents' source population. For migrant individuals with parents coming from different source populations, we sampled one allele from the source population of each parent. Finally, samples were generated by drawing n_i individu als from each local population, keeping track of the ancestry of both alleles for each sampled individual.

Implementation details: Each MCMC was run for $1,030,000$ iterations. The first $20,000$ iterations consist of short pilot runs used to tune up the proposal distributions to obtain acceptance rates between 25 and 45%. The next $10,000$ iterations were discarded as burn-in and the remaining observations were sampled every 100 iterations, giving a sample size of $10,000$ for each analysis.

To take into account model uncertainty, parameters are estimated using Bayesian model-averaging methods. The only exception to this rule are the regression parameters, which are model specific and, therefore, were estimated using the subset of values corresponding to the model with the highest posterior probability. Finally, posterior model probabilities are obtained by observing the number of times the chain visits each alternative model.

Posterior estimates are based on the sample mean except for the deviation from

the regression σ^2, which usually has a highly asymmetric posterior distribution. In this latter case we used the posterior mode, which was estimated using kernel density estimation.

We investigated the effect of varying three parameters: the level of genetic differentiation, $F_{ST} = \{0.01, 0.05, 0.10, 0.25\}$; the proportion of nonmigrant alleles, $m_{ii} = \{0.7, 0.9\}$; and the number of populations, $I = \{4, 6\}$.

For each parameter set we generated 10 independent genetic data sets as described above. The results we present below are averages across these 10 replicates. As a measure of accuracy we also present the relative mean square errors (RMSE).

Results: We investigated the performance of our method to provide reliable estimates under different scenarios of migration and genetic differentiation and number of populations studied. We consider first the effects on model determination, and then we address the influence on migration rate estimates and finally on individual assignments.

When the immigration rate is high ($m_{ii} = 0.7$; see Table 7.1), estimates of posterior model probabilities are strongly influenced by the degree of genetic differentiation (F_{ST}). When differentiation is low ($F_{ST} = 0.01$), the method fails to identify the model used to generate the synthetic data. However, the correct model is identified when $F_{ST} > 0.01$, and, moreover, its posterior model probability increases steadily with increasing genetic differentiation. The estimation of regression parameters is also influenced by the magnitude of F_{ST} but to a lesser degree. The RMSE decreases with increasing genetic differentiation but the bias is largely unaffected. Thus, it is the accuracy of the estimates (as illustrated by the HPDIs) that is influenced by F_{ST}. The proportion of the variance that remains unexplained by the model, σ^2, decreases as genetic differentiation increases.

Decreasing the immigration rate ($m_{ii} = 0.90$) has a detrimental effect on estimates (Table 7.2). Although the true model is correctly identified for $FST > 0.01$, its posterior probability is lower than that observed when $m_{ii} = 0.70$. Estimates of regression parameters are more biased and less accurate (wider HPDIs), leading to higher RMSEs. Also, the proportion of the variance that remains unexplained, σ^2, is larger. Note, however, that as was the case before, the quality of all estimates improves with increasing genetic differentiation.

Increasing the number of populations studied ($I = 6$) improves model determination (Table 7.3). More precisely, the posterior probability of the true model is strongly increased and the proportion of variance that remains unexplained decreases sharply (see Table 7.3 and last columns of Tables 7.1 and 7.2). However, the effect on the quality of the regression parameter estimates is somewhat decreased since the bias and the RMSE increase. Nevertheless, the width of the HPDIs decreases, indicating that the precision increases.

Posterior estimates for various levels of genetic differentiation and high gene flow.

		F_{ST}			
Factors included		0.01	0.05	0.10	0.25
none		**0.621**	0.073	0.026	0.015
$G1$		0.142	0.061	0.028	0.016
$G2$		0.170	0.184	0.082	0.052
***G*1 and *G*2**		0.045	**0.472**	**0.637**	**0.701**
with interaction		0.022	0.210	0.227	0.216
Parameter	True value	Estimate/RMSE/95% HPDI			
α_1	-0.900	[a]	-0.921	-0.974	-0.945
			0.057	0.027	0.003
			[-1.754;-0.097]	[-1.639;-0.324]	[-1.503;-0.370]
α_2	1.100	[a]	1.137	1.178	1.149
			0.013	0.011	0.002
			[0.285;2.068]	[0.504;1.869]	[0.554;1.745]
σ^2	–	0.389	0.426	0.355	0.306
		–	–	–	–
		[0.120;1.159]	[0.121;2.858]	[0.106;2.010]	[0.107;1.578]
Assignments					
Missassignments		0.754	0.280	0.110	0.002
Probabilities[b]		0.223	0.700	0.883	0.996

Table 7.1: Posterior model probabilities, regression parameter mean estimates and assignment accuracy for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ are shown.

[a] The regression parameter is not included in the model with the highest posterior model probability.

[b] Maximum posterior assignment probabilities averaged across all individuals.

Posterior estimates for various levels of genetic differentiation and low gene flow.

		F_{ST}			
Factors included		0.01	0.05	0.10	0.25
none		**0.444**	0.106	0.056	0.028
$G1$		0.122	0.072	0.055	0.030
$G2$		0.291	0.301	0.122	0.073
***G*1 and *G*2**		0.075	**0.323**	**0.521**	**0.638**
with interaction		0.068	0.197	0.246	0.231
Parameter	True value	Estimate/RMSE/95% HPDI			
α_1	-0.900	[a]	-0.858	-1.122	-1.010
			0.123	0.101	0.015
			[-1.984;0.263]	[-2.091;-0.191]	[-1.723;-0.321]
α_2	1.100	[a]	1.403	1.314	1.173
			0.127	0.071	0.008
			[0.197;2.687]	[0.387;2.279]	[0.473;1.886]
σ^2	–	0.486	0.513	0.452	0.352
		–	–	–	–
		[0.131;3.067]	[0.125;4.117]	[0.120;3.090]	[0.110;1.956]
Assignments					
Missassignments		0.808	0.134	0.046	0.002
Probabilities[b]		0.288	0.847	0.946	0.997

Table 7.2: Posterior model probabilities, regression parameter mean estimates and assignment accuracy for synthetic data generated with porportions of nonmigrant alleles set to $m_{ii} = 0.90$ are shown.

[a] The regression parameter is not included in the model with the highest posterior model probability.

[b] Maximum posterior assignment probabilities averaged across all individuals.

Posterior estimates for scenario with six populations.

		$F_{ST} = 0.25$	
Factors included		$m_{ii} = 0.70$	$m_{ii} = 0.90$
none		0.000	0.000
$G1$		0.000	0.000
$G2$		0.000	0.001
***G*1 and *G*2**		**0.915**	**0.883**
with interaction		0.085	0.116
Parameter	True value	Estimate/RMSE/95%HPDI	
α_1	-0.900	-1.056	-1.022
		0.031	0.022
		[-1.416;-0.707]	[-1.521;-0.537]
α_2	1.100	1.244	1.246
		0.018	0.018
		[0.960;1.531]	[0.856;1.659]
σ^2	–	0.164	0.213
		–	–
		[0.080;0.381]	[0.096;0.566]
Assignments			
Missassignments		0.010	0.002
Probabilities[a]		0.985	0.997

Table 7.3: Model determination, regression parameters, mean estimates and assignment accuracy for synthetic data when varying nonmigrant gene proportions m_{ii}.

[a] The regression parameter is not included in the model with the highest posterior model probability.

Estimates of gene migration rates improve with increasing genetic differentiation (Table 7.4). The bias decreases sharply between $F_{ST} = 0.01$ and 0.1 and then remains very low. Note that the only cases where the HPDI does not include the true value correspond to the case with the weakest genetic differentiation ($F_{ST} = 0.01$). When the number of nonmigrant genes decreases we observe the same pattern but in this case the number of estimates for which the HPDIs do not include the true value is smaller and corresponds only to the estimates of nonmigrant proportions (Table 7.5).

In terms of posterior individual assignments, increasing genetic differentiation improves the quality of the estimation (see the bottom three rows of Tables 7.1 and 7.2). That is, the proportion of individuals that are misassigned decreases while the average posterior assignment probability increases. Decreasing the proportion of migrant genes also improves the quality of assignments; the proportion of misassignments decreases and the average posterior probabilities with which individuals are assigned increase. The effect of varying the number of populations is very small, being somewhat more distinguishable when the proportion of nonmigrants is larger (see bottom three rows of Tables 7.1-7.3).

We also investigated what is the effect of using explicative variables that are different from the ones used to generate the synthetic data. The results (Table 7.6) show that the highest posterior probability is assigned to the null model, which indicates that the method does not wrongly identify as important factors that are not responsible for the observed migration pattern.

It is also important to investigate the effect of the amount of data used for the estimation, which can be characterized by the sample sizes and number of loci scored. The effect of decreasing the sample size from 100 to 50 individuals per population does not have much of an effect on posterior model probabilities while estimates of regression parameters have a slightly larger bias and wider HPDIs leading to somewhat larger RMSEs (compare last column of Table 7.1 with Table 7.7). Migration rate estimates show no increase in bias but their HPDIs are larger (compare Tables 7.8 and 7.9). Finally, the quality of the assignments is barely influenced by a decrease in the sample sizes (compare Tables 7.1 and 7.7). The effect of increasing the number of loci scored from 10 to 20 does not have an effect on model determination, estimates of regression parameters, and migration rates when the level of genetic differentiation is moderate ($F_{ST} = 0.10$) (results not shown). The only result that changes is the proportion of individuals that are misassigned, which decreases from 0.002 to 0. We also carried out analysis of a scenario with $F_{ST} = 0.05$ and in this case, increasing the number of loci from 10 to 20 decreased the width of the HPDIs for migration rate estimates and improved the accuracy of individual assignments.

Migration estimates for various levels of genetic differentiation and high migration rate.

		F_{ST}			
Migration rate	True value	0.01	0.05	0.10	0.25
m_{11}	0.700	0.355[a]	0.663	0.702	0.710
m_{12}	0.023	0.173	0.036	0.021	0.020
m_{13}	0.003	0.222[a]	0.007	0.002	0.002
m_{14}	0.274	0.250	0.293	0.275	0.269
m_{21}	0.018	0.184	0.027	0.021	0.020
m_{22}	0.700	0.359[a]	0.656	0.720	0.710
m_{23}	0.227	0.252	0.240	0.213	0.220
m_{24}	0.055	0.205	0.076	0.047	0.050
m_{31}	0.259	0.224	0.291	0.273	0.256
m_{32}	0.028	0.252[a]	0.029	0.027	0.025
m_{33}	0.700	0.336[a]	0.662	0.689	0.704
m_{34}	0.013	0.188[a]	0.020	0.010	0.014
m_{41}	0.149	0.243	0.151	0.136	0.142
m_{42}	0.081	0.211	0.111	0.085	0.079
m_{43}	0.070	0.185	0.074	0.061	0.064
m_{44}	0.700	0.362[a]	0.664	0.719	0.715

Table 7.4: Posterior estimates averaged across analyses of 10 simulated data sets with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ are shown.

[a] The 95% HPDI does not contain the true value.

Migration estimates for various levels of genetic differentiation and high migration rate.

		F_{ST}			
Migration rate	True value	0.01	0.05	0.10	0.25
m_{11}	0.900	0.407[a]	0.867	0.910	0.910
m_{12}	0.008	0.232	0.014	0.006	0.006
m_{13}	0.001	0.182	0.004	0.001	0.001
m_{14}	0.091	0.179	0.115	0.083	0.084
m_{21}	0.006	0.209	0.008	0.005	0.005
m_{22}	0.900	0.428[a]	0.890	0.901	0.905
m_{23}	0.076	0.197	0.088	0.078	0.075
m_{24}	0.018	0.165	0.014	0.016	0.014
m_{31}	0.086	0.244	0.098	0.087	0.086
m_{32}	0.009	0.170	0.009	0.009	0.010
m_{33}	0.900	0.449[a]	0.882	0.901	0.900
m_{34}	0.004	0.137	0.012	0.004	0.005
m_{41}	0.050	0.165	0.046	0.042	0.043
m_{42}	0.027	0.181	0.033	0.025	0.023
m_{43}	0.023	0.181	0.056	0.020	0.019
m_{44}	0.900	0.473[a]	0.865	0.913	0.915

Table 7.5: Posterior estimates averaged across analyses of 10 simulated data sets with proportions of nonmigrant alleles set to $m_{ii} = 0.90$ are shown.

[a] The 95% HPDI does not contain the true value.

Posterior model estimates when testing for nonexplanatory factors.

	F_{ST}			
Factors included	0.01	0.05	0.10	0.25
none	**0.636**	**0.533**	**0.512**	**0.505**
$G1$	0.127	0.224	0.243	0.241
$G2$	0.184	0.157	0.151	0.158
***G*1 and *G*2**	0.039	0.065	0.071	0.076
with interaction	0.013	0.022	0.022	0.021
Assignments				
Missassignments	0.741	0.282	0.110	0.002
Probabilities	0.218	0.694	0.881	0.996

Table 7.6: Posterior model probabilities and assignment accuracy when varying level of genetic differentiation F_{ST} and testing for two nonexplanatory factors (i.e. different from the ones we used for generating migration rates) are shown. Data were simulated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$.

Model estimates when sampling 50 individuals per population.

Factors included		Posterior probabilities
none		0.021
$G1$		0.021
$G2$		0.074
***G*1 and *G*2**		**0.651**
with interaction		0.233
Parameter	True value	Estimate/RMSE/95%HPDI
α_1	-0.900	-0.960
		0.008
		[-1.608;-0.322]
α_2	1.100	1.208
		0.014
		[0.536;1.909]
σ^2	–	0.349
		[0.108;1.922]
Assignments		
Missassignments		0.008
Probabilities		0.989

Table 7.7: Posterior model probabilities, regression parameter mean estimates and assignment accuracy for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ and level of genetic differentiation set to $F_{ST} = 0.25$ are shown.

Migration estimates when sampling 50 individuals per population.

Parameter	True value	Posterior estimate	RMSE	95% HPDI
m_{11}	0.700	0.702	< 0.001	[0.612;0.790]
m_{12}	0.023	0.026	0.043	[0.003;0.056]
m_{13}	0.003	0.002	0.070	[0.000;0.010]
m_{14}	0.274	0.270	0.001	[0.185;0.356]
m_{21}	0.018	0.011	0.149	[0.000;0.029]
m_{22}	0.700	0.722	0.001	[0.632;0.808]
m_{23}	0.227	0.220	0.001	[0.142;0.301]
m_{24}	0.055	0.047	0.025	[0.013;0.086]
m_{31}	0.259	0.238	0.007	[0.158;0.322]
m_{32}	0.028	0.026	0.030	[0.003;0.055]
m_{33}	0.700	0.725	0.001	[0.637;0.811]
m_{34}	0.013	0.011	0.033	[0.000;0.028]
m_{41}	0.149	0.136	0.008	[0.074;0.201]
m_{42}	0.081	0.075	0.006	[0.031;0.124]
m_{43}	0.070	0.067	0.001	[0.025;0.115]
m_{44}	0.700	0.721	0.001	[0.634;0.808]

Table 7.8: Estimates based on the posterior mean, RMSE and 95% HPDI are reported for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ and the level of genetic differentiation set to $F_{ST} = 0.25$.

Migration rate estimates when sampling 100 individuals per population.

Parameter	True value	Posterior estimate	RMSE	95% HPDI
m_{11}	0.700	0.710	< 0.001	[0.647;0.772]
m_{12}	0.023	0.020	0.018	[0.004;0.038]
m_{13}	0.003	0.002	0.218	[0.000;0.006]
m_{14}	0.274	0.269	< 0.001	[0.208;0.323]
m_{21}	0.018	0.020	0.014	[0.005;0.038]
m_{22}	0.700	0.710	< 0.001	[0.647;0.772]
m_{23}	0.227	0.220	0.001	[0.165;0.277]
m_{24}	0.055	0.050	0.009	[0.023;0.079]
m_{31}	0.259	0.256	< 0.001	[0.198;0.317]
m_{32}	0.028	0.025	0.009	[0.007;0.046]
m_{33}	0.700	0.704	< 0.001	[0.640;0.766]
m_{34}	0.013	0.014	0.013	[0.002;0.029]
m_{41}	0.149	0.142	0.002	[0.097;0.190]
m_{42}	0.081	0.079	0.001	[0.045;0.115]
m_{43}	0.070	0.064	0.007	[0.033;0.098]
m_{44}	0.700	0.715	< 0.001	[0.652;0.777]

Table 7.9: Estimates based on the posterior mean, RMSE and 95% HPDI are reported for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ and the level of genetic differentiation set to $F_{ST} = 0.25$.

7.3 Application to real data

We use the human genome diversity cell line panel Centre d'Etude du Polymorphisme-Humain (HGDP-CEPH) presented by Cann *et al.* (2002) to illustrate how our method can be used to make inferences about the factors influencing migration patterns. In our example we selected a subset of eight populations, all from Pakistan (see Figure 7.2), corresponding to 200 individuals (25 per population). We grouped together the Balochi and Brahui samples because the STRUCTURE analyses carried out by Rosenberg *et al.* (2002) place them in the same genetic cluster (see their Figure 2). Also, instead of using all 377 loci we did a first screening using an improved version of Beaumont and Balding's (2004) method to identify outlier loci that could be influenced by selection. On the basis of this screening we selected a total of 247 loci that were used in the analysis.

The effect of distance is supposed to be one of the main factors in determining gene flow in many species, but other factors such as altitude can influence geographic isolation and, therefore, migration patterns. We use our method to evaluate the relative importance of these two factors. We obtained pairwise geographic distances from latitude and longitude coordinates and also calculated the difference in altitude between each focal population and all other populations. Cann *et al.* (2002) give the geographic coordinates of each population as sample intervals; thus we used the gravity center of the area for the calculation of geographic distances between populations. With two parameters we can define five alternative models, which are presented in Table 7.10.

As was the case for the simulation study we used short pilot runs to tune up the proposal distributions to achieve reasonable acceptance ratios. To ensure convergence we increased the burn-in to 10^6 and the sample size to $20,000$ and used a thinning interval of 50 iterations. Some of the population-specific F_{ST} values are < 0.01 (see Table 7.11), the level of genetic differentiation that our simulation study identified as problematic for the estimation of parameters. This example, therefore, provides us with an opportunity to illustrate the problems that may arise when our method (or any other MCMC-based method) is used in scenarios with weak genetic differentiation. In these situations, it is necessary to run many independent replicates and compare their results; in the present case we used 10 runs. In 6 of them, the most probable model included altitude only and in all cases there was a posterior probability of at least 50%. The second most probable model included both factors. However, in 4 other runs two other models, one including distance only and the other including both distance and altitude, gave similar high posterior probabilities while the model including altitude only was ranked third. Given these results, we followed Faubet *et al.* (2007) and chose the run with the lowest deviance for estimation purposes. The Bayesian deviance has been proposed as a measure of model fit by a number of authors

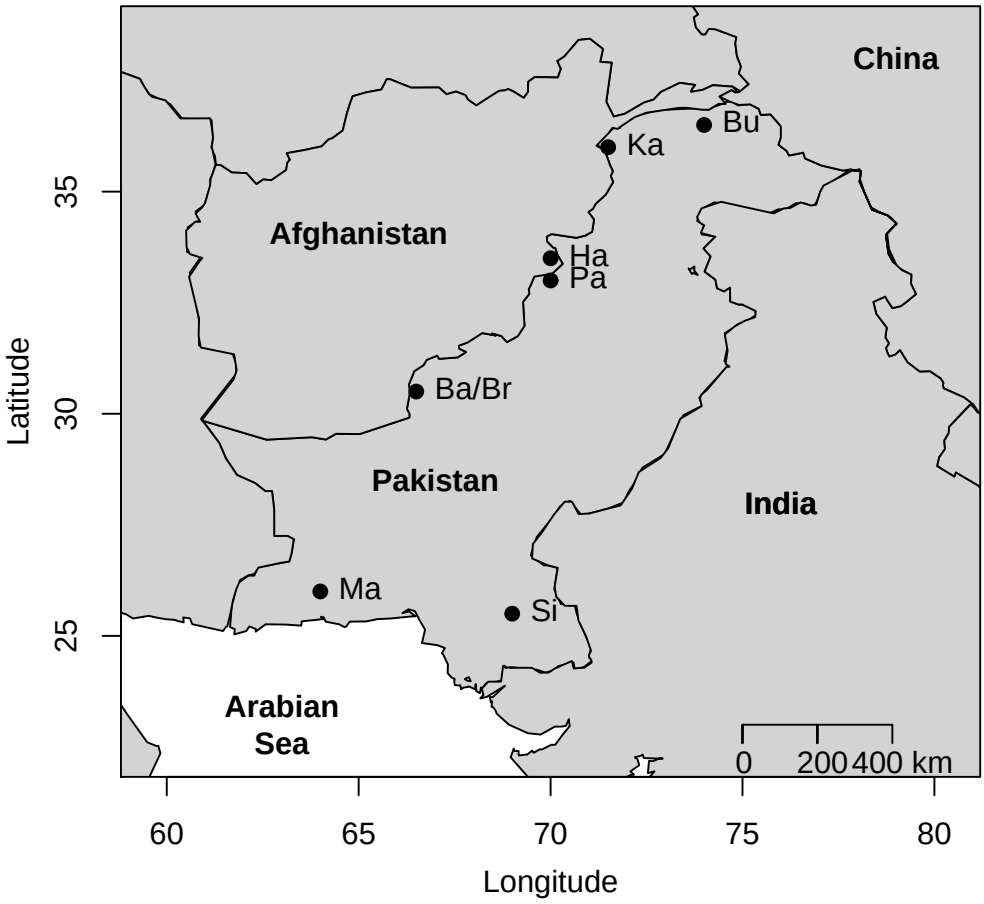


Figure 7.2: Geographic locations of sampled populations. Solid circles represent centers of gravity of sampled areas of Pakistan. Abbreviations for population names are as follows: Ba, Balochi; Br, Brahui; Bu, Burusho; Ha, Hazara; Ka, Kalash; Ma, Makrani; Pa, Pathan; Si, Sindhi.

Posterior model probabilities for Pakistani human data set.

		Estimate/95%HPDI		
Factors included	$\Pr(\mathcal{M}\vert\dots)$	α_1	α_2	α_3
none	0.064			
Distance	0.093	-0.756		
		[-2.88;0.417]		
Altitude	**0.550**		-1.74	
			[-3.35;0.338]	
Distance and Altitude	0.232	-0.469	-1.55	
		[-2.95;0.639]	[-2.91;0.747]	
with interaction	0.061	-0.524	-1.64	0.302
		[-2.66;0.71]	[-3.4;0.599]	[-1.2;1.44]

Table 7.10: Posterior model probabilities for the human data set when considering geographic distances and differences in altitudes as environmental factors. Posterior estimates for regression parameters are based on the mode and 95% HPDI. The maximum a posteriori estimate of σ^2 is 0.657 with 95% HPDI ranging from 0.089 to 11.1.

(Faubet *et al.* 2007 and references therein) and in our specific case we considered the assignment component of the total deviance, $D_{assign} = -2 \log \Pr(\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{m})$. Table 7.10 presents these results. The model with the highest posterior probability includes altitude only and the second most probable model includes both altitude and distance. In this latter case, the regression coefficients for the effects of altitude and distance are both negative, indicating that, as expected, both factors decrease migration rates between populations. Note, however, that the former seems to have a stronger effect (i.e. larger absolute value).

Table 7.12 presents the mode and HPDI of migration rates between populations. Although the sum of maximum posterior estimates does not necessarily add up to one, we used them as estimators because of the inherent asymmetry of migration rate posterior distributions. There are three populations that do not receive migrants (Burusho, Bu; Hazara, Ha; and Kalash, Ka) and they correspond to those located in high-altitude areas. Moreover, two of these populations (Bu and Ka) do not seem to send migrants either and a third one (Ha) seems to contribute very little to the gene pool of the Pathan. The population with the highest proportion of migrant genes is the Sindhi, which receives migrants mainly from Balochi/Brahui (Ba/Br). Three other populations have a somewhat lower proportion of migrant genes (Ba/Br; Makrani, Ma; and Pathan, Pa). In the case of Ba/Br, most of the genes come from Ma, and, conversely, most of the genes of Ma come from Ba/Br. Finally, the Pathan receive similar proportions of genes from Ba/Br, Ha, and Sindhi (Si). In general, there are frequent gene exchanges among southern populations while northern populations remain fairly isolated. The best explanation for this migration pattern is altitude differences with the most isolated populations being at high altitude and the least isolated ones at low altitude.

Finally, the mean and mode of inbreeding coefficient estimates are somewhat large when compared to F_{ST} estimates but this is not the case if we compare the lower bounds of the HPDIs (Table 11). Still, there are three local populations (Ba/Br, Ma, and Pa) for which the lower bound of F_{IS} HPDIs is > 0.04 while that of F_{ST}s is much lower. A potential explanation for this result could be that samples were taken from adult individuals and, therefore, the data set does not fit model assumptions concerning the moment at which sampling takes place. However, we do not have information concerning the age group involved in the sampling.

7.4 Discussion

We present a new method for the estimation of recent migration rates that also allows for making inferences about the factors that influence gene flow in subdivided populations. It focuses on the F1 descendants of migrant individuals and, therefore, estimates the probability that a given individual migrated during the

Local population F_{ST}s and inbreeding coefficient for Pakistani populations.

	Mean/Mode/95%HPDI	
Population	F_{ST}	F
Ba/Br	0.010	0.091
	0.010	0.099
	$[0.006; 0.015]$	$[0.042; 0.132]$
Bu	0.009	0.010
	0.009	0.009
	$[0.005; 0.013]$	$[10^{-5}; 0.033]$
Ha	0.015	0.0155
	0.015	0.0156
	$[0.010; 0.020]$	$[10^{-5}; 0.040]$
Ka	0.048	0.014
	0.049	0.0133
	$[0.040; 0.058]$	$[10^{-6}; 0.037]$
Ma	0.007	0.116
	0.007	0.105
	$[0.001; 0.016]$	$[0.065; 0.195]$
Pa	0.009	0.0917
	0.008	0.0904
	$[0.003; 0.015]$	$[0.046; 0.142]$
Si	0.017	0.0579
	0.016	0.0586
	$[0.009; 0.028]$	$[0.005; 0.125]$

Table 7.11: Estimates are based on posterior mean and mode.

Migration rates between Pakistani populations.

	Mean/Mode/95%HPDI						
From/Into	Ba/Br	Bu	Ha	Ka	Ma	Pa	Si
Ba/Br	0.690	0.000	0.000	0.000	0.220	0.060	0.280
	0.670	0.000	0.000	0.000	0.150	0.020	0.300
	$[0.427; 0.900]$	$[10^{-15}; 10^{-8}]$	$[10^{-9}; 10^{-7}]$	$[10^{-10}; 10^{-7}]$	$[0.008; 0.638]$	$[10^{-8}; 0.339]$	$[0.021; 0.668]$
Bu	0.000	1.000	0.000	0.000	0.010	0.01	0.010
	0.000	1.000	0.000	0.000	0.000	0.000	0.000
	$[0.004; 0.086]$	$[1.000; 1.000]$	$[10^{-9}; 10^{-7}]$	$[10^{-10}; 10^{-7}]$	$[0.001; 0.190]$	$[10^{-4}; 0.207]$	$[10^{-12}; 0.189]$
Ha	0.010	0.000	1.000	0.000	0.010	0.040	0.010
	0.000	0.000	1.000	0.000	0.000	0.020	0.000
	$[0.001; 0.135]$	$[10^{-16}; 10^{-8}]$	$[1.000; 1.000]$	$[10^{-14}; 10^{-7}]$	$[10^{-12}; 0.168]$	$[0.010; 0.246]$	$[10^{-12}; 0.151]$
Ka	0.000	0.000	0.000	1.000	0.000	0.010	0.000
	0.000	0.000	0.000	1.000	0.000	0.000	0.000
	$[10^{-4}; 0.073]$	$[10^{-10}; 10^{-8}]$	$[10^{-15}; 10^{-7}]$	$[1.000; 1.000]$	$[10^{-4}; 0.079]$	$[10^{-9}; 0.128]$	$[10^{-4}; 0.132]$
Ma	0.220	0.000	0.000	0.000	0.680	0.060	0.150
	0.230	0.000	0.000	0.000	0.740	0.010	0.030
	$[0.014; 0.488]$	$[10^{-15}; 10^{-8}]$	$[10^{-15}; 10^{-7}]$	$[10^{-10}; 10^{-7}]$	$[0.233; 0.935]$	$[10^{-9}; 0.451]$	$[10^{-9}; 0.560]$
Pa	0.010	0.000	0.000	0.000	0.030	0.740	0.020
	0.000	0.000	0.000	0.000	0.000	0.760	0.000
	$[0.003; 0.143]$	$[10^{-15}; 10^{-8}]$	$[10^{-9}; 10^{-7}]$	$[10^{-9}; 10^{-7}]$	$[0.001; 0.302]$	$[0.362; 0.936]$	$[0.003; 0.251]$
Si	0.070	0.000	0.000	0.000	0.050	0.070	0.530
	0.060	0.000	0.000	0.000	0.030	0.050	0.520
	$[0.009; 0.225]$	$[10^{-15}; 10^{-8}]$	$[10^{-14}; 10^{-7}]$	$[10^{-9}; 10^{-7}]$	$[0.001; 0.265]$	$[10^{-9}; 0.339]$	$[0.145; 0.874]$

Table 7.12: Estimates are based on posterior mean and mode.

previous generation. Our approach also estimates various other population-specific parameters such as local F_{ST}, inbreeding coefficients, and allele frequencies. The method requires data from codominant markers such as RFLPs, microsatellites, allozymes, and SNPs and environmental data specific to each local population. Note, however, that the modeling of dispersal barriers (mountains, roads, deforested areas, etc.) between pairs of populations can be introduced by considering landscape resistance measures usually used by landscape ecologists (see, e.g., McRae 2006).

We generated synthetic data following the inference model described above to investigate the effect of varying levels of genetic differentiation, proportions of non-migrant genes, and numbers of populations, loci, and individuals. The results of this simulation study indicate that the method can provide reliable estimates when global F_{ST} values are $> 1\%$, the number of loci is only 10, and sample sizes are of the order of 50 individuals per population. Additionally, the identification of the environmental factors influencing migration is easier when migration rates are high and the number of local populations considered increases. We did not investigate the effect of varying the degree of polymorphism (i.e. the number of allelic classes) or the effect of unsampled populations. We expect that increasing polymorphism will increase accuracy while the effect of unobserved populations is more likely to decrease it depending on true migration rates between unsampled and sampled populations. Our simulation study could be extended to take into account these considerations. Additionally, it would be desirable to consider demographic scenarios that differ from the one assumed by the inference model to test the robustness of our method.

We applied our method to a previously published microsatellite human data set for which local F_{ST}s are within the range of values that our simulation study identified as problematic for parameter estimation. As expected, we observed convergence problems for this application and followed the approach of Faubet *et al.* (2007) to minimize them (see previous section for a more detailed explanation). We found that altitude influences recent migration among Pakistani populations and that gene exchanges are more frequent in the south than in the north of Pakistan. Geographic distance seems to have little effect on migration, a result that can be explained by the limited geographic scale considered and the fact that even in poorly developed areas there are many means of transportation that facilitate movement of humans. On the other hand, altitude can represent an important barrier particularly in winter when populations at high altitude can remain isolated for long periods of time.

The estimation of migration rates has proved to be a very difficult task. Several methods exist for this purpose; some of them estimate long-term migration rates and are based on coalescent theory (e.g., MIGRATE, Beerli and Felsenstein 2001) while others provide recent migration rate estimates and are based on multilo-

cus genotype approaches (e.g., BAYESASS, Wilson and Rannala 2003). All recent methods for estimating migration rates rely on MCMC approaches and require one to pay special attention to convergence issues (Faubet *et al.* 2007). This is particularly important when genetic differentiation among populations is weak. This caveat also applies to our method, and the human example we present illustrates how to deal with these problems.

Being a multilocus genotype approach, our method resembles in many respects BAYESASS. It is important to note, however, that this resemblance is only superficial because we do not assume the same sampling scheme and we allow for high migration rates. Indeed, as opposed to Wilson and Rannala (2003) we assume that sampling takes place after reproduction and before migration. This was done to avoid the low migration rate restriction underlying their method and to allow migration rates to vary between 0 and 1. More specifically, Wilson and Rannala's (2003) formulation provides estimates of migration rates restricted to the interval $(0, 1/3)$ and assumes that m is very small because to account for individuals with mixed ancestry (i.e. individuals whose alleles come from two different populations) they need to consider individuals that arrived one generation before sampling takes place. Thus, they are forced to assume that at most half of an individual's alleles comes from another population. In our case, we do not have this restriction because after reproduction the alleles of a given individual can come from any population. Doing this, however, precludes us from distinguishing between first-generation and second-generation migrants. Nevertheless, we can consider cases where parents are migrants from two different populations while BAYESASS considers only a single migrant ancestor.

The information used by our estimation method is the gametic disequilibrium generated by migration, which increases as genetic differentiation among local populations increases. Indeed, limited migration is very effective in increasing differentiation of gamete types among the subpopulations by random genetic drift (Ohta 1982). The strength of this gametic disequilibrium can be measured through the genotype of migrant individuals (or descendants from recent migrants) or through the gamete haplotype frequencies. Clearly, the former corresponds to short-term migration while the latter corresponds to the effect of long-term migration. All this implies that if the long-term migration is very high, the signature left by recent migration events will be weak. In the case of our method, the simulation study indicates that reliable estimates can be obtained when the effective number of migrants is less than five (i.e. $F_{ST} \geq 0.05$). The gametic disequilibrium due to long-term migration can also lead to a deviation from the hypothesis of independence among loci used to derive the likelihood function. This is a problem shared by all the methods that estimate migration rates from multilocus genotype data. The potential biases that could be introduced due to this problem require

a very detailed simulation study, using an individually based model that produces synthetic data that allow for the estimation of gametic disequilibrium.

Another improvement introduced in our method is the use of the F-model first proposed by Balding and Nichols (1995). This feature allows us to take into account the population admixture that may have taken place before the last generation of migration. Additionally, as pointed out by Falush *et al.* (2003), the implementation of this model permits identification of subtle population subdivisions and, therefore, improves the estimation of allele-frequency distributions when genetic differentiation is weak. This in turn improves the estimation of migration rates as shown by a pilot study comparing the performance of our method with and without the F-model (results not shown). All the improvements implemented by our method lead to good mixing properties of the MCMC and therefore minimize convergence problems. We stress, however, that users should always carefully check the convergence of the MCMC by running multiple analyses and comparing their results.

An important feature of our method is that besides simply estimating migration rates it also identifies the factors that influence them. We use the same approach as that first proposed by Gaggiotti *et al.* (2004), which consists of using a Dirichlet prior for the immigration rates and linking its shape parameters with the environmental data, using a generalized linear model. In the present case, however, we do not consider models without the constant factor (i.e. the regression intercept). This was done because our experience with the application of this type of method (Gaggiotti *et al.* 2004; Foll and Gaggiotti 2006) indicates that models excluding this parameter almost always had null posterior probabilities. These results can be explained by the fact that the regression intercept captures the effects of factors that act at a larger geographic scale than that considered for the metapopulation under study. It also takes into account behavioral characteristics of the species under study that remain the same regardless of the environment. In fact, the regression intercept influences only the variance of immigration rates, which increases as α_0 decreases. For example, we expect that the variance of the migration rate between two given populations will be larger for species that can disperse very long distances than for species with very poor dispersal abilities. In this case, then we expect to obtain estimates of the intercept that are smaller for the former.

In our approach we assumed that the probability of observing nonmigrant alleles in any given population is independent of environmental factors. The underlying rationale for this is that local environmental conditions will influence only emigration rates but do not have any effect on the immigration rates that are the focus of our estimation method. Ideally we would also like to estimate emigration rates. As Wilson and Rannala (2003) point out, this could be done if we know local population sizes or, alternatively, if we could develop a method that

can make use of temporal samples. However, such approaches are likely to involve much more complex likelihood functions that will necessarily lead to a worsening of convergence problems that are typical of complex methods that use MCMC approaches.

The software that implements the method incorporates features that facilitate the interpretation of results. For example, it provides estimates of both means and modes, which allows the user to choose the best parameter estimator depending on the shape of the posterior distribution (which is also provided by the software). Indeed, when posterior distributions are asymmetric, posterior estimates based on the mode and on the mean are rather different and the former provides a better way of describing the results. Thus, users should always have a look at the shape of posterior distributions to choose appropriate estimators.

Bayesian methods such as the one we present here are powerful tools for the study of natural populations. Users, however, should keep in mind that their application requires some expertise on the computational methods underlying their implementation, particularly on MCMC approaches. These issues are discussed more in detail in Faubet *et al.* (2007) and also in the user manuals of several of the currently available methods. If these recommendations are followed, population biologists will be able to extract highly valuable information about the species under study.

7.5 Acknowledgements

Most of the computations presented in this paper were performed on the cluster HealthPhy (CIMENT, Grenoble). We are grateful to Matthieu Foll for providing us the software to identify outlier loci in the human dataset. We also thank Olivier François and two anaymous reviewers for their useful suggestions that helped to improve the manuscript. This work was supported by the Fond National de la Science (grant ACI-Impbio-2004-42-ADGP). P.F. holds a Ph.D. studentship from the Ministère de la Recherche. The software implementing the method is available for all platforms at http://www-leca.ujf-grenoble.fr/logiciels.htm.

7.6 Appendix: prior distributions for parameters

We take the following priors for each parameter discussed in the text.

Probability to observe nonmigrant genes: We assume that nonmigrant proportions are not influenced by environmental factors and therefore use a uniform distribution:

$$m_{ii} \backsim \mathcal{U}(0,1), \text{ i.e. } f(m_{ii}) = \begin{cases} 1 & \text{if } m_{ii} \in (0,1) \\ 0 & \text{otherwise} \end{cases} \tag{7.12}$$

Probability to observe migrant genes: We use a Dirichlet prior for the rate of migrant genes contributed by local populations other than the focal one, $m^{\star}_{ij}$,

$$\boldsymbol{m}^{\star}{}_i|\boldsymbol{\psi}_i \backsim Dir(\boldsymbol{\psi}_i), \text{ i.e. } f(\boldsymbol{m}^{\star}{}_i|\boldsymbol{\psi}_i) = \Gamma\left(\sum_{j\neq i}\psi_{ij}\right)\prod_{j\neq i}\frac{m_{ij}^{\star\psi_{ij}-1}}{\Gamma(\psi_{ij})} \tag{7.13}$$

where the $m^{\star}_{ij}$s are given by Equation (7.6).

Shape parameters for the Dirichlet prior: As the ψ_{ij}s must be positive we use a log-normal distribution,

$$\log\psi_{ij}|\boldsymbol{\alpha},\sigma^2,\boldsymbol{G} \backsim \mathcal{N}(\mu_{ji},\sigma^2), \text{ i.e. } f(\psi_{ij}|\boldsymbol{\alpha},\sigma^2,\boldsymbol{G}) = \frac{1}{\psi_{ij}\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\log\psi_{ij}-\mu_{ij})^2}{2\sigma^2}\right) \tag{7.14}$$

where μ_{ij} is given by the regression (7.8).

Regression coefficients: We use a normal distribution,

$$\alpha \backsim \mathcal{N}(0,\sigma_\alpha^2), \text{ i.e. } f(\alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{\alpha^2}{2\sigma_\alpha^2}\right) \tag{7.15}$$

where $\sigma_\alpha^2 = 10$.

Deviation from the regression: We assume that σ^2 has an inverse-gamma distribution:

$$\tau = \sigma^{-2} \backsim Gamma(a_\tau, b_\tau), \text{ i.e. } f(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\tau^{a_\tau-1}\exp(-\tau b_\tau) \tag{7.16}$$

where $a_\tau, b_\tau = 1$

Local population F_{ST}s: As θ_is must be positive we use a log-normal distribution:

$$\log \theta_i \backsim \mathcal{N}(\omega, \xi), \text{ i.e. } f(\theta_i) = \frac{1}{\theta_i \sqrt{2\pi\xi}} \exp\left(\frac{(\log \theta_i - \omega)^2}{2\xi}\right) \tag{7.17}$$

where $\omega = \xi = 1$.

Metapopulation allele frequencies: We use an uninformative Dirichlet prior:

$$\tilde{\boldsymbol{p}}_l \backsim Dir(\lambda, \ldots, \lambda), \text{ i.e. } f(\tilde{\boldsymbol{p}}_l) = \frac{\Gamma(k_l\lambda)}{\Gamma(\lambda)^{k_l}} \prod_{a=1}^{k_l} \tilde{p}_{la}^{\lambda-1} \tag{7.18}$$

where k_l is the number of alleles observed at locus l in the metapopulation and $\lambda = 1$.

Population alleles frequencies: We use a Dirichlet prior:

$$\boldsymbol{p}_{il}|\theta_i, \tilde{\boldsymbol{p}}_l \backsim Dir(\theta_i \tilde{\boldsymbol{p}}_l), \text{ i.e. } f(\boldsymbol{p}_{il}|\theta_i, \tilde{\boldsymbol{p}}_l) = \Gamma(\theta_i) \prod_{a=1}^{k_l} \frac{p_{ila}^{\theta_i \tilde{p}_{la}-1}}{\Gamma(\theta_i \tilde{p}_{la})} \tag{7.19}$$

Population specific inbreeding coefficients: We use a uniform distribution:

$$F_i \backsim \mathcal{U}(-1, 1), \text{ i.e. } f(F_i) = \begin{cases} 1/2 & \text{if } F_i \in (-1, 1) \\ 0 & \text{otherwise} \end{cases} \tag{7.20}$$

Ancestry assignments: Following Wilson and Rannala (2003), we use a multinomial prior:

$$\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{m}_i \backsim Mult(n_i, \boldsymbol{m}_i), \text{ i.e. } Pr(\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{m}_i) = n_i! \prod_{j \leq k} \frac{\widetilde{m}_{ijk}^{n_{ijk}}}{n_{ijk}!} \tag{7.21}$$

where n_{ijk} is the number of individuals sampled from population i that belongs to ancestry class (j, k) and $\widetilde{m}_{ijk}$s is given by Equation (7.1).

7.7 Supplementary Information: Details of the MCMC methods

As we want to estimate both posterior model probabilities and corresponding parameters, we require two transition steps in our MCMC scheme. First posterior distributions of parameters under a given model are updated (within-model moves). Then we propose to change from the current model to another (between-model moves).

Within-model moves Parameters of model $\mathcal{M}$ are updated following the transition kernel described below. We use either Metropolis-Hastings moves or Gibbs sampling.

Updating $\boldsymbol{m}$: This move consists of two steps. First we update proportions of non-migrant genes and then migrant gene rates. Let $\boldsymbol{m}'$ be the new proposed value for the migration matrix $\boldsymbol{m}$ with elements m_{ij} and $m_{ij'}$ replaced by m'_{ij} and $m'_{ij'}$ respectively. Corresponding values for $\widetilde{m}_{ijk}$s are calculated using equation (7.1).
Non migrant gene proportions: For each population i, we propose a new value m'_{ii} for non migrant proportion setting

$$m'_{ii} = u \tag{7.22}$$

where $u \backsim \mathcal{U}(\max(0, m_{ii} - e_{nm}), \min(1, m_{ii} + e_{nm}))$ and e_{nm} is some incremental value. Then we adjust immigration rates so that they add up to one, i.e. for each migrant source $j \neq i$

$$m'_{ij} = m_{ij} \times \frac{1 - m'_{ii}}{1 - m_{ii}} \tag{7.23}$$

This move is accepted with probability $\alpha(\boldsymbol{m}, \boldsymbol{m}') = \min(1, A)$ where

$$A = \frac{Pr(\boldsymbol{M}|\boldsymbol{m}')f(\boldsymbol{m}'|\boldsymbol{\psi})q(\boldsymbol{m}, \boldsymbol{m}')}{Pr(\boldsymbol{M}|\boldsymbol{m})f(\boldsymbol{m}|\boldsymbol{\psi})q(\boldsymbol{m}', \boldsymbol{m})} \tag{7.24}$$

Proportion of migrant genes: Updating immigration rates for each population i, denoted as vector $\boldsymbol{m}_i$, is performed pairwise because the elements must sum to one. We randomly choose two populations that differ from the focal one, say populations j and j'. Then we propose a new immigration rate into population i from population j by setting

$$m'_{ij} = u \tag{7.25}$$

where $u \backsim \mathcal{U}(\max(0, m_{ij} - e_m), \min(m_{ij} + m_{ij'}, m_{ij} + e_m))$ and e_m is some incremental value. Finally we set

$$m'_{ij'} = m_{ij} + m_{ij'} - m'_{ij} \tag{7.26}$$

as the elements of the $\boldsymbol{m}'_i$ must add up to one.

This move is accepted with probability $\alpha(\boldsymbol{m}, \boldsymbol{m}') = \min(1, A)$ where

$$A = \frac{Pr(\boldsymbol{M}|\boldsymbol{m}')f(\boldsymbol{m}'|\boldsymbol{\psi})q(\boldsymbol{m}, \boldsymbol{m}')}{Pr(\boldsymbol{M}|\boldsymbol{m})f(\boldsymbol{m}|\boldsymbol{\psi})q(\boldsymbol{m}', \boldsymbol{m})} \tag{7.27}$$

Updating $\boldsymbol{\psi}$: Gene flow intensities (i.e. shape parameters of the prior for migration rate) are updated individually. Let $\boldsymbol{\psi}'$ be the matrix $\boldsymbol{\psi}$ with element ψ_{ij} replaced by ψ'_{ij} where i and j are two distinct populations. For each of them, we set

$$\log \psi'_{ij} = u \tag{7.28}$$

where $u \backsim \mathcal{N}(\log \psi_{ij}, \sigma^2_\psi)$ and σ^2_ψ is some incremental value.

This move is accepted with probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \min(1, A)$ where

$$A = \frac{f(\boldsymbol{m}|\boldsymbol{\psi}')f(\boldsymbol{\psi}'|\boldsymbol{\alpha}, \sigma^2, \boldsymbol{G})q(\boldsymbol{\psi}, \boldsymbol{\psi}')}{f(\boldsymbol{m}|\boldsymbol{\psi})f(\boldsymbol{\psi}|\boldsymbol{\alpha}, \sigma^2, \boldsymbol{G})q(\boldsymbol{\psi}', \boldsymbol{\psi})} \tag{7.29}$$

Updating $\boldsymbol{\alpha}$: This update is performed using a multivariate move with the Gibbs sampler. Let $Z_{\mathcal{M}}$ be the matrix whose $I(I-1)$ rows are given by

$$\left(1 \,\middle|\, G_{ij}^{(1)} \quad \dots \quad G_{ij}^{(R)} \,\middle|\, G_{ij}^{(1)}G_{ij}^{(2)} \quad \dots \quad G_{ij}^{(1)}G_{ij}^{(R)} \,\middle|\, \dots \,\middle|\, G_{ij}^{(R-1)}G_{ij}^{(R)} \right)$$

for each pair of distinct populations $i \neq j$. Here the subscript $\mathcal{M}$ denotes the model. Let $\boldsymbol{\alpha}_{\mathcal{M}}$ be the column vector with all regression parameters in the model. Here it includes all environmental factors with corresponding first-order interactions. If another model is considered, corresponding $Z_{\mathcal{M}}$ matrix (resp. $\boldsymbol{\alpha}_{\mathcal{M}}$ vector) is obtained by removing appropriate columns (resp. rows). The posterior conditional distribution for the vector $\boldsymbol{\alpha}_{\mathcal{M}}$ is given by

$$f(\boldsymbol{\alpha}_{\mathcal{M}}|\dots) \propto f(\boldsymbol{\psi}|\boldsymbol{\alpha}_{\mathcal{M}}, \sigma^2, \boldsymbol{G})f(\boldsymbol{\alpha}_{\mathcal{M}})$$

According to the prior for regression parameters (7.15), $\boldsymbol{\alpha}_{\mathcal{M}} \backsim \mathcal{N}(\boldsymbol{0}, \sigma^2_\alpha I_{|\mathcal{M}|})$ where $I_{|\mathcal{M}|}$ is the identity matrix of size $|\mathcal{M}|$, the number of parameters in the model.

Following the matrix construction above, we construct the column vector $\log \boldsymbol{\psi}$ ($I(I-1)$ rows). According to equation (7.7), $\log \boldsymbol{\psi} \backsim \mathcal{N}(Z_{\mathcal{M}}\boldsymbol{\alpha}_{\mathcal{M}}, \sigma_2 I_{I(I-1)})$, thus

$$\begin{aligned} f(\boldsymbol{\alpha}_{\mathcal{M}}|\ldots) &\propto \exp\left(-\frac{1}{2\sigma^2}(\log\boldsymbol{\psi} - Z_{\mathcal{M}}\boldsymbol{\alpha}_{\mathcal{M}})^T(\log\boldsymbol{\psi} - Z_{\mathcal{M}}\boldsymbol{\alpha}_{\mathcal{M}})\right)\exp\left(-\frac{1}{2\sigma_\alpha^2}\boldsymbol{\alpha}_{\mathcal{M}}^T\boldsymbol{\alpha}_{\mathcal{M}}\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\alpha}_{\mathcal{M}}^T\left\{\frac{1}{\sigma^2}Z_{\mathcal{M}}^T Z_{\mathcal{M}} + \frac{1}{\sigma_\alpha^2}I_{|\mathcal{M}|}\right\}\boldsymbol{\alpha}_{\mathcal{M}} - \frac{2}{\sigma^2}\boldsymbol{\alpha}_{\mathcal{M}}^T Z_{\mathcal{M}}^T\log\boldsymbol{\psi}\right]\right) \end{aligned}$$

Hence $\boldsymbol{\alpha}_{\mathcal{M}}|\ldots \backsim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{M}}, \Sigma_{\mathcal{M}})$ where

$$\Sigma_{\mathcal{M}}^{-1} = \frac{1}{\sigma^2}Z_{\mathcal{M}}^T Z_{\mathcal{M}} + \frac{1}{\sigma_\alpha^2}I_{|\mathcal{M}|} \text{ and } \boldsymbol{\mu}_{\mathcal{M}} = \frac{1}{\sigma^2}\Sigma_{\mathcal{M}}Z_{\mathcal{M}}^T\log\boldsymbol{\psi} \tag{7.30}$$

Updating σ^2: This update is also performed using the Gibbs sampler. For notational convinience let τ be σ^{-2}. We can write the distribution of τ given all other parameters applying Bayes' rule. According to equations (7.14) and (7.16) we have

$$\begin{aligned} f(\tau|\ldots) &\propto f(\boldsymbol{\psi}|\boldsymbol{\alpha}, \tau, \boldsymbol{G}) \times f(\tau) \\ &\propto \tau^{I(I-1)/2}\exp\left(-\frac{\tau}{2}\sum_{i\neq j}(\log\psi_{ij} - \mu_{ij})^2\right) \times \tau^{a_\tau - 1}\exp\left(-\tau b_\tau\right) \\ &\propto \tau^{(a_\tau + I(I-1)/2)-1}\exp\left(-\tau(b_\tau + S/2)\right) \end{aligned} \tag{7.31}$$

where $S = \sum_{i\neq j}(\log\psi_{ij} - \mu_{ij})^2$. Then τ given all other parameters is gamma distributed with shape parameter $a_\tau + I^2/2$ and rate parameter $b_\tau + S/2$, i.e.

$$\tau|\ldots \backsim G(a_\tau + I(I-1)/2, b_\tau + S/2) \tag{7.32}$$

Updating $\boldsymbol{\theta}$: Local population F_{ST}s are updated individually for each population i. Let $\boldsymbol{\theta}'$ be the vector $\boldsymbol{\theta}$ with element θ_i replaced by θ_i'. For each population q we set

$$\log\theta_i' = u \tag{7.33}$$

where $u \backsim \mathcal{N}(\log\theta_i, \sigma_\theta^2)$ and σ_θ^2 is some incremental value.

This move is accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min(1, A)$ where

$$A = \frac{f(\boldsymbol{p}|\tilde{\boldsymbol{p}}, \boldsymbol{\theta}')f(\boldsymbol{\theta}')q(\boldsymbol{\theta}, \boldsymbol{\theta}')}{f(\boldsymbol{p}|\tilde{\boldsymbol{p}}, \boldsymbol{\theta})f(\boldsymbol{\theta})q(\boldsymbol{\theta}', \boldsymbol{\theta})} \tag{7.34}$$

Updating $\tilde{\boldsymbol{p}}$: Updating allele frequencies of the metapopulation for each locus l, denoted as vector $\tilde{\boldsymbol{p}}$, is performed pairwise because its elements must add up to one. Let $\tilde{\boldsymbol{p}}'$ be the population allele frequencies $\tilde{\boldsymbol{p}}$ with elements $\tilde{p}_{la}$ and $\tilde{p}_{la'}$ respectivly replaced by $\tilde{p}'_{la}$ and $\tilde{p}'_{la'}$ where a and a' are two randomly choosen alleles observed at locus l. We propose a new frequency for allele a by setting

$$\tilde{p}'_{la} = u \tag{7.35}$$

where $u \backsim \mathcal{U}(\max(0, \tilde{p}_{la} - e_p), \min(\tilde{p}_{la} + \tilde{p}_{la'}, \tilde{p}_{la} + e_p))$ and e_p is some incremental value. Finally, we set

$$\tilde{p}'_{la'} = \tilde{p}_{la} + \tilde{p}_{la'} - \tilde{p}'_{la} \tag{7.36}$$

as the elements of $\tilde{\boldsymbol{p}}'$ must add up to one.

This move is accepted with probability $\alpha(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{p}}') = \min(1, A)$ where

$$A = \frac{f(\boldsymbol{p}|\tilde{\boldsymbol{p}}', \boldsymbol{\theta}) f(\tilde{\boldsymbol{p}}') q(\boldsymbol{p}, \boldsymbol{p}')}{f(\boldsymbol{p}|\tilde{\boldsymbol{p}}, \boldsymbol{\theta}) f(\tilde{\boldsymbol{p}}) q(\boldsymbol{p}', \boldsymbol{p})} \tag{7.37}$$

Updating $\boldsymbol{p}$: Updating allele frequencies of each population i at locus l, denoted as vector $\boldsymbol{p}_{il}$, is performed pairwise because its elements must add up to one. Let $\boldsymbol{p}'_{il}$ be the population allele frequencies $\boldsymbol{p}_{il}$ with elements p_{ila} and $p_{ila'}$ respectivly replaced by p'_{ila} and $p'_{ila'}$ where a and a' are two randomly choosen alleles observed at locus l. We propose a new frequency for allele a setting

$$p'_{ila} = u \tag{7.38}$$

where $u \backsim \mathcal{U}(\max(p_{min}, p_{ila} - e_p), \min(p_{ila} + p_{ila'} - p_{min}, p_{ila} + e_p))$, $p_{min} = \max(0, F_i/(F_i - 1))$ is the minimal value for population i allele frequencies (because of inbreeding coefficient constraint) and e_p is some incremental value. Finally, we set

$$p'_{ila'} = p_{ila} + p_{ila'} - p'_{ila} \tag{7.39}$$

as the elements of $\boldsymbol{p}'_{il}$ must add up to one.

This move is accepted with probability $\alpha(\boldsymbol{p}, \boldsymbol{p}') = \min(1, A)$ where

$$A = \frac{Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}') f(\boldsymbol{p}'|\tilde{\boldsymbol{p}}, \boldsymbol{\theta}) q(\boldsymbol{p}, \boldsymbol{p}')}{Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}) f(\boldsymbol{p}|\tilde{\boldsymbol{p}}, \boldsymbol{\theta}) q(\boldsymbol{p}', \boldsymbol{p})} \tag{7.40}$$

Note that only individuals with either allele a or allele a' copy at locus l originating from population i are concerned by this move.

When the F-model is not used, shape parameters for the Dirichlet prior, $\theta_i \tilde{p}_{la}$s, are replaced by λ (i.e. uninformative prior).

Updating $\boldsymbol{F}$: Inbreeding coefficients $\boldsymbol{F}$ are updated for each population i. Let $\boldsymbol{F}'$ be the vector of inbreeding coefficient $\boldsymbol{F}$ with element F_i replaced by F_i'. Let p_{min} denote the minimum of allele frequencies across loci in population i. Because we consider inbreeding coefficient, its value is constrained to be greater than $F_{min} = \max(-1, -p_{min}/(1-p_{min}))$. So we propose a new inbreeding coefficient for population i by setting

$$F_i' = u \tag{7.41}$$

where $u \backsim \mathcal{U}(\max(F_{min}, F_i - e_F), \min(1, F_i + e_F))$ and e_F is some incremental value.

This move is accepted with probability $\alpha(\boldsymbol{F}, \boldsymbol{F}') = \min(1, A)$ where

$$A = \frac{Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{F}', \boldsymbol{p}) f(\boldsymbol{F}') q(\boldsymbol{F}, \boldsymbol{F}')}{Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}) f(\boldsymbol{F}) q(\boldsymbol{F}', \boldsymbol{F})} \tag{7.42}$$

Note that only individuals whose both parents come from population i are concerned by this move.

Updating $\boldsymbol{M}$: Assignments are updated for each individual h. Let $\boldsymbol{M}'$ be the ancestry state $\boldsymbol{M}$ with element $\boldsymbol{M}_h$ replaced by $\boldsymbol{M}_h'$. Individual h is sampled from population S_h. We propose a new ancestry (i', j') in population S_h uniformly among the $I^2 - 1$ states other than its current state (i, j).

This move is accepted with probability $\alpha(\boldsymbol{M}, \boldsymbol{M}') = \min(1, A)$ where

$$A = \frac{Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}', \boldsymbol{F}, \boldsymbol{p}) Pr(\boldsymbol{M}'|\boldsymbol{m}) q(\boldsymbol{M}, \boldsymbol{M}')}{Pr(\boldsymbol{X}|\boldsymbol{S}; \boldsymbol{M}, \boldsymbol{F}, \boldsymbol{p}) Pr(\boldsymbol{M}|\boldsymbol{m}) q(\boldsymbol{M}', \boldsymbol{M}, \boldsymbol{t})} \tag{7.43}$$

Between-model moves We want to identify which environmental factors have influenced the migration process. Thus we have to compare different competing models depending on which factors are included in the regression (7.8). Alternative models can be expressed by canceling appropriate αs. Reversible jump MCMC (Geyer 1995) provide a natural extension to MCMC allowing changes in the dimension of the state vector $\boldsymbol{\alpha}$. Using this method we can estimate posterior probabilities of the alternative models and their parameters.

In order to move from model $\mathcal{M}$ to model $\mathcal{M}'$, we generate $\boldsymbol{\alpha}'$ from some proposal distribution $q(.)$ independent of $\boldsymbol{\alpha}$ and $\mathcal{M}$. This move is accepted with probability $\alpha(\mathcal{M}, \mathcal{M}') = \min(1, A)$ where

$$A = \frac{f(\boldsymbol{\alpha}', \mathcal{M}'| \ldots) q(\boldsymbol{\alpha}) Pr(\mathcal{M}' \to \mathcal{M})}{f(\boldsymbol{\alpha}, \mathcal{M}| \ldots) q(\boldsymbol{\alpha}') Pr(\mathcal{M} \to \mathcal{M}')} |J| \tag{7.44}$$

As we choose $\mathcal{M}'$ uniformly at random from all possible models,

$$\frac{Pr(\mathcal{M}' \to \mathcal{M})}{Pr(\mathcal{M} \to \mathcal{M}')} = 1$$

We drop the current $\boldsymbol{\alpha}$ vector and generate a new $\boldsymbol{\alpha}'$ one for the proposed model. Thus, the Jacobian reduces to the identity matrix with determinant 1. Using Bayes' rule, posterior probability of model and corresponding parameters is given by

$$f(\boldsymbol{\alpha}, \mathcal{M}| \ldots) \propto f(\boldsymbol{\psi}|\boldsymbol{\alpha}_{\mathcal{M}}, \sigma^2, \boldsymbol{G}) f(\boldsymbol{\alpha}_{\mathcal{M}}) Pr(\mathcal{M})$$

We assume no prior information on alternative models so that $Pr(\mathcal{M}) \propto 1$ and the corresponding terms cancel in the acceptance ratio.

Brooks *et al.* (2003) showed that the best choice for the proposal distribution $q(.)$ is to take its full conditional distribution of $\boldsymbol{\alpha}$ under model $\mathcal{M}$, i.e. the multivariate normal distribution described for regressor updates $f_{\boldsymbol{\mu},\Sigma}$.

Finally, the acceptance ratio of between-model moves is given by

$$A = \frac{f(\boldsymbol{\psi}|\boldsymbol{\alpha}_{\mathcal{M}'}, \sigma^2, \boldsymbol{G}) f(\boldsymbol{\alpha}_{\mathcal{M}'}) f_{\boldsymbol{\mu}_{\mathcal{M}},\Sigma_{\mathcal{M}}}(\boldsymbol{\alpha}_{\mathcal{M}})}{f(\boldsymbol{\psi}|\boldsymbol{\alpha}_{\mathcal{M}}, \sigma^2, \boldsymbol{G}) f(\boldsymbol{\alpha}_{\mathcal{M}}) f_{\boldsymbol{\mu}_{\mathcal{M}'},\Sigma_{\mathcal{M}'}}(\boldsymbol{\alpha}_{\mathcal{M}'})} \tag{7.45}$$

Troisième partie

L'influence des facteurs environnementaux sur l'adaptation locale

Chapitre 8

Sélection naturelle et adaptation locale

L'étude des mécanismes d'adaptation d'une espèce à son environnement est fondamentale en biologie de la conservation et en biologie évolutive. Au niveau génétique, la démarche consiste à (i) d'identifier les loci soumis à la sélection pour, ensuite, (ii) déterminer les facteurs environnementaux responsables de l'adaptation locale. La génétique des populations permet d'aborder ces problèmes et fournit d'importantes informations pour l'amélioration et la gestion des espèces, pour l'identification de gènes de résistance à des pathogènes ou impliqués dans le déterminisme de maladies génétiques.

8.1 La sélection naturelle

La sélection naturelle explique l'adaptation des espèces à leur milieu selon s'il favorise ou pénalise les porteurs de certains traits. Ce processus évolutif agit à travers la survie et/ou le succès reproducteur des individus, en fonction de leur capacité dans un environnement donné. Ainsi, pour observer la sélection d'un caractère dans une population, il faut que ce caractère varie d'un individu à un autre, que cette variation soit héritable et qu'elle ait un effet sur la survie ou la fécondité des individus.

Les évenements de mutation et de recombinaison génétique sont à l'origine de la variabilité des phénotypes, ils sont la source de la diversité génétique. Dans la mesure où ces processus agissent sur les allèles, ils touchent directement le matériel héréditaire. Selon l'environnement, un allèle pourra être avantageux, délétaire ou neutre et conditionnera ainsi la survie ou la fécondité des individus qui le portent. De cette manière, la sélection naturelle contribue à faire apparaître des structures génétiques contrastées dans des milieux différents et augmente la différentiation

génétique.

La sélection naturelle agit sur les fréquences alléliques selon des scénarios complexes, dans l'espace mais aussi dans le temps. En effet, l'environnement change selon les époques, les allèles les mieux adaptés en un lieu aujourd'hui diffèrent probablement de ceux d'hier ou de demain. Ainsi, les changements environnementaux contribuent à redistribuer la diversité génétique dans l'espace et influencent l'histoire évolutive des espèces.

Lorsqu'une espèce occupe un habitat qui présente des gradients environnementaux, des variations des fréquences alléliques peuvent apparaître sous l'action de la migration et de la sélection naturelle. D'une part, la migration (force évolutive neutre) redistribue et maintient la diversité génétique. D'autre part, selon l'environnement, la sélection naturelle (cf. §8.2.3) peut conduire soit à la fixation des allèles favorables - et à l'élimination des allèles désavantageux - (sélection directionnelle), soit au maintien du polymorphisme lorsqu'il est avantageux d'être hétérozygote (sélection balancée). Ainsi, la structure spatiale de la diversité génétique dépend des effets de compensation entre migration et sélection naturelle. Par exemple, si un allèle est défavorisé par la sélection naturelle dans une population, il peut y être maintenu s'il est sans cesse réintroduit par la migration.

Selon les effets de compensations entre migration et sélection naturelle, il y a différentes conséquences sur la diversité génétique : si la migration est plus importante que la sélection ou s'il y équilibre entre les deux forces, les allèles, même s'ils sont défavorisés, seront maintenus ; si la migration est plus faible que la sélection, les allèles désavantageux seront éliminés. Ainsi migration et sélection peuvent contribuer à l'apparition de structures génétiques contrastées dans différents environnements.

8.2 Les modèles de la génétique des populations

Pour prendre en compte de l'effet de la sélection naturelle, la génétique des populations utilisent les notions de valeurs sélectives et de coefficients de sélection. Dans les modèles, ces paramètres gouvernent l'évolution des fréquences alléliques et permettent de décrire différents types de sélections.

8.2.1 Notion de valeur sélective, coefficient de sélection

La grandeur qui mesure la degré d'adaptation d'un individu à son environnement est la valeur sélective [1] de son génotype, notée w. Cette grandeur correspond au nombre de descendants que les individus porteurs de ce génotype peuvent espérer laisser à la génération suivante.

1. appelée aussi fitness.

La valeur sélective d'un génotype intègre deux composantes : la viabilité, v, et la fécondité, f, des individus porteurs du génotype en question. La viabilité d'un génotype mesure la probabilité qu'un individu ayant ce génotype arrive à l'âge reproducteur ; la fécondité quantifie le nombre moyen de descendants qu'il produira. Selon ces définitions, la valeur sélective d'un génotype est

$$W = fv \tag{8.1}$$

soit le produit de la fécondité et de la viabilité.

Ainsi, la quantité W, la *fitness absolue*, mesure le nombre de descendants qu'un individu peut espérer laisser lors de la génération suivante en fonction de son génotype. Généralement, la performance d'un génotype est exprimée de façon relative à un génotype de référence (e.g. le mieux adapté) et implique le calcul de la *fitness relative*. Si les valeurs sélectives absolues d'un génotype et d'un autre de référence sont respectivement W et W_{ref}, alors la fitness relative, w est

$$w = \frac{W}{W_{ref}} \tag{8.2}$$

Si le génotype de référence est le plus performant des génotypes observés, alors les valeurs sélectives relatives sont comprises entre 0 et 1.

La différence entre la fitness relative, w, et 1 s'appelle le *coefficient de sélection*, noté s ; cette grandeur mesure le taux de réduction ou d'accroissement de la fitness de chaque génotype par rapport au génotype de référence.

La fitness ou le coefficient de sélection d'un génotype sont des grandeurs difficiles à mesurer directement car ils peuvent varier à travers le génome (phénomènes de compensation) et selon l'environnement (qui lui aussi peut changer dans le temps). Cependant, les modèles de la génétique des populations permettent d'étudier l'évolution des fréquences alléliques sous l'effet de la sélection.

8.2.2 Evolution d'une population sous l'effet de la sélection

Le cas le plus simple est celui d'un locus biallélique dans une population d'effectif infini (i.e. pas de dérive). Les générations sont supposées non chevauchantes et les croisements aléatoires. Il n'y a ni migration, ni mutation ; la sélection agit sur la viabilité mais pas sur la fécondité des individus. Les deux allèles observés - A et a, de fréquences respectives p et $q = 1 - p$ - donnent trois génotypes - AA, Aa et aa - dont les valeurs sélectives sont respectivement w_{AA}, w_{Aa} et w_{aa}.

Le tableau 8.1 décrit l'évolution de la structure génotypique sous les hypothèses mentionnées ci-dessus. Les fréquences génotypiques de la génération suivante dépendent de la valeur sélective des génotypes et de $\bar{w} = w_{AA}p^2 + 2w_{Aa}pq + w_{aa}q^2$, la valeur sélective moyenne de la population.

Génotype	AA	Aa	aa
Fréquence avant sélection	p^2	$2pq$	q^2
Valeur sélective	w_{AA}	w_{Aa}	w_{aa}
Après sélection	$p^2\frac{w_{AA}}{\bar{w}}$	$2pq\frac{w_{AA}}{\bar{w}}$	$q^2\frac{w_{aa}}{\bar{w}}$

TABLE 8.1 – Evolution des fréquences génotypiques sous l'effet de la sélection.

Connaissant la structure génotypique, il est possible de déterminer la structure allélique pour obtenir la fréquence de l'allèle A lors de la prochaine génération, i.e.

$$p' = p\frac{pw_{AA} + qw_{Aa}}{\bar{w}} \tag{8.3}$$

et les variations de cette même fréquence

$$\delta p = pq\frac{(w_{AA} - w_{Aa})p + (w_{Aa} - w_{aa})q}{\bar{w}} \tag{8.4}$$

Le devenir des fréquences alléliques dépend du signe de δp qui dépend lui même des valeurs sélectives. Si δp est positif alors la fréquence de l'allèle va augmenter, s'il est nul alors il y a équilibre, s'il est négatif alors la fréquence de l'allèle va diminuer. Selon les différentes relations d'ordre entre les valeurs 'sélective, les variations des fréquences alléliques conduiront à différents types de sélection.

8.2.3 Différents types de sélection

L'analyse des différentes relations d'ordre entre valeurs sélective permet d'identifier trois modes de sélection :

- la sélection directionnelle en faveur d'un allèle : lorsque $w_{AA} \geq w_{Aa} \geq w_{aa}$ (ou le cas symétrique $w_{AA} \leq w_{Aa} \leq w_{aa}$), la fréquence de l'allèle A (ou a) va augmenter jusqu'à ce que l'allèle soit fixé. Elle conduit à la fixation de l'allèle avantageux et à l'élimination de l'allèle défavorable.
- la superdominance ou la sélection en faveur de l'hétérozygote : lorsque $w_{AA} \leq w_{Aa}$ et $w_{Aa} \geq w_{aa}$, elle maintient le polymorphisme (effet balancé), même les allèles létaux.
- la sous dominance ou la sélection contre l'hétérozygote : lorsque $w_{AA} \geq w_{Aa}$ et $w_{Aa} \leq w_{aa}$, elle conduit à des équilibres instables et joue probablement un rôle important dans la spéciation

Des scenarios plus complexes, dans lesquels interviennent plusieurs processus évolutifs, permettent de prédire l'évolution des fréquences alléliques. Dans leur application ces modèles ont pour but de détecter les loci sous sélection et d'identifier les facteurs qui influencent ce processus.

8.3 Détecter et mesurer la sélection naturelle

L'un des enjeux de la génétique des populations est (i) d'identifier les gènes soumis à la sélection pour, ensuite, (ii) étudier les facteurs responsables de l'adaptation

locale. Différentes méthodes ont été développées pour répondre à la première question (e.g. Vitalis et al. 2001 ; Beaumont and Balding 2004) ; elles reposent sur l'identification de loci au comportement atypique[2] par rapport à une évolution neutre (i.e. en l'absence de sélection). Cependant, comme l'a souligné Gaggiotti (2006), peu de méthodes permettent de répondre à la deuxième question qui constitue un problème difficile. L'approche de Novembre et al. (2005) pour étudier la mutation $\Delta 32$ chez l'homme constitue un premier pas qui mérite d'être suivi.

Les méthodes qui cherchent à identifier les loci sous sélection utilisent des criblages génomiques pour détecter des gènes qui n'auraient pas eu une évolution neutre. L'idée de base est qu'un locus sous sélection directionnelle sera plus différentié qu'un locus neutre alors qu'un locus sous sélection balancée le sera moins. Ainsi le niveau de différentiation génétique permet de distinguer des loci au comportement atypique d'un bruit de fond neutre. La principale difficulté de ces méthodes est la possible confusion entre sélection naturelle et histoire démographique qui peut conduire à la détection de faux positifs.

Une fois qu'un gène sous sélection est détecté, il s'agit d'identifier les facteurs responsables de la sélection naturelle. L'approche introduite de Novembre et al. (2005) consiste à corréler la distribution spatiale des fréquences alléliques du locus étudié à des gradients environnementaux. Leur idée était de simuler l'évolution de la fréquence d'un gène pour retrouver la distribution spatiale observée. Le modèle utilisé pour les simulations, qui intègre les effets de la migration et de la sélection, repose sur la théorie des clines génétiques introduite ci-dessous.

Chez de nombreuses espèces les fréquences alléliques varient le long de gradients environnementaux ; de tels phénomènes sont appelés *clines génétiques*. A partir de cette observation, les généticiens des populations ont élaboré des modèles pour expliquer l'apparition de clines. Les premiers développements théoriques furent ceux de Fisher (1937) qui proposa un modèle mathématique[3] pour étudier la diffusion des gènes sous l'effet de la migration et de la sélection naturelle. Ses travaux furent suivis par ceux de Haldane (1948) qui considéra des coefficients de sélection qui varient dans l'espace pour mesurer l'intensité de la sélection naturelle. Par la suite les travaux de Slatkin and Maruyama (1974), de Felsenstein (1975) et de Nagylaki (1978) ont permis d'étendre le modèle de Fisher en considérant l'effet de la dérive génétique. D'autres travaux théoriques (Slatkin 1973 ; Nagylaki 1976) ont permis d'explorer plus en détails le modèle de Fisher, mais les applications de ces modèles restent peu nombreuses (Gaggiotti 2006).

Dans les modèles de clines avec migration et sélection, l'évolution spatio-temporelle de la fréquence d'un gène est modélisée par une équation aux dérivées partielles (edp) : la migration intervient à travers un terme de diffusion spatiale (dérivée

2. outlier
3. Fisher's wave of advance

seconde en espace, $\frac{\partial^2 p}{\partial x^2}$) alors que la sélection naturelle apparaît dans un terme de réaction (fonction de croissance locale, $sf(p)$). Les paramètres du modèle sont le coefficient de diffusion D, qui mesure le déplacement quadratique moyen (migration), le coefficient de sélection s et le mode de sélection f. Ainsi, lorsque la migration est homogène, les variations temporelles de la fréquence p d'un allèle est égale à la somme des effets de la migration et de la sélection naturelle :

$$\frac{\partial p}{\partial t} = \underbrace{D\frac{\partial^2 p}{\partial x^2}}_{\text{migration}} + \underbrace{sf(p)}_{\text{sélection}} \tag{8.5}$$

Dans le cas général, l'équation (8.5) ci-dessus n'a pas de solutions analytiques, seules des méthodes numériques permettent de la résoudre. D'autre part, pour que le problème soit bien posé, il est nécessaire de donner la distribution initiale de la fréquence et les conditions aux limites de l'habitat (déterminées selon des considérations biologiques).

La figure 8.1 présente la solution numérique de l'équation (8.5) lorsque la fréquence de l'allèle est supposée uniforme, égale à 1/2, à l'instant initial et qu'il n'y a pas de flux de gènes aux bornes de l'habitat. Le coefficient de sélection dépend linéairement de la position de telle sorte que l'allèle est défavorisé à gauche du domaine et favorisé à droite et la croissance locale logistique (i.e. $f(p) = p(1-p)/\bar{w}$, modèle de sélection additif). Au cours du temps, (i.e. $t \nearrow$) la migration et la sélection font apparaître un cline pour atteindre une situation d''équilibre.

Les méthodes numériques de résolution de l'edp (8.5) permettent de générer des clines selon différentes valeurs du coefficient de sélection. Intégrées dans un schéma ABC, les méthodes de simulation des clines serviront à estimer des paramètres liés à la sélection naturelle.

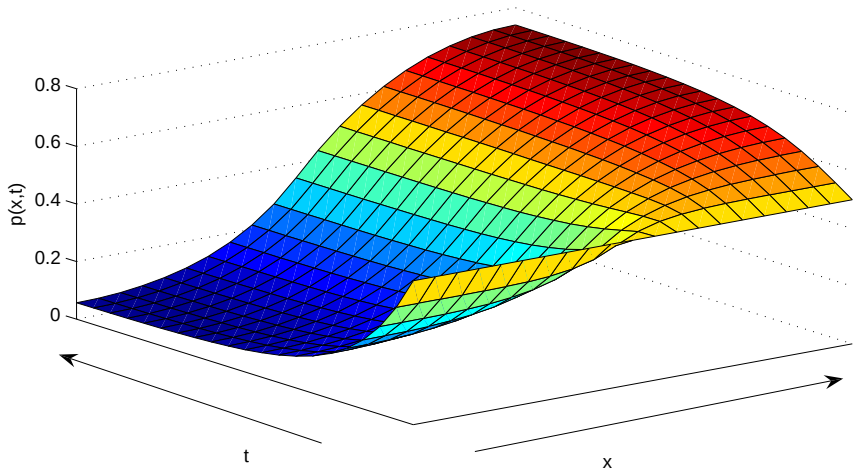


FIGURE 8.1 – Evolution de la fréquence d'un allèle sous l'effet de la migration et de la sélection naturelle. La migration est homogène dans l'espace et constante dans le temps. La pression de sélection est fonction de la position mais constante dans le temps. La variable d'espace est x, t représente le temps et $p(x,t)$ la fréquence de l'allèle à la position x au temps t.

Chapitre 9

Article III

Ce chapitre présente la méthode proposée par Faubet et al. (in prep.) pour estimer le coefficient de sélection le long d'un gradient latitudinal et mesurer l'influence de la latitude sur l'adaptation locale. Les paramètres d'un modèle bayésien qui repose sur la théories des clines génétiques sont estimés par une méthode ABC. Une étude de sensibilité, à partir de données simulées, est présentée ainsi qu'une application à des données réelles chez *Fundulus heteroclitus*.

Chapter 10

An ABC method for the study of genetic clines: combining environmental and genetic data for the inference of dispersal and selection parameters

Pierre Faubet, Jean-Louis Martiel and Oscar E. Gaggiotti

We present a new ABC approach for the estimation of dispersal and selection parameters. Our method estimates the variance of the parent-offspring dispersal distribution, the heterozygous effect and the influence of environmental factors on selection coefficients. Our ABC algorithm relies on the simulation of genetic clines under migration and selection at a single biallelic locus. We carry out a sensitivity analysis using synthetic data to evaluate our method. Our results show that we can estimates dispersal and selection parameter accurately when the environmental gradient is well established. We also demonstrate our ABC approach by analyzing *Fundulus heteroclitus* allozyme data along a latitudinal gradient. Our results suggest that our method is sensitive to the noise in observed data.

10.1 Introduction

Many natural populations exhibit a directional change of a character and/or gene frequency over geographical distance, a spatial pattern for which Huxley (1939) coined the term 'cline'. It is generally accepted that clines are maintained by a balance between dispersal and selection. Although in some cases dispersal may be negligible and selection may maintain a stable equilibrium at each locality along a transect (Barton and Hewitt 1985), in most cases the homogenising effect of migration is important and limits the extent of differentiation among populations at the selected locus. There is a large body of theory devoted to the study of the conditions under which a cline can be maintained by the balance between migration and selection. Although some spatially discrete models of selection and migration have been used to study clines, (e.g. Moody 1979, 1981; Nagylaki 1977), most of the studies have used the more tractable continuous counterparts (e.g. Slatkin 1973; Nagylaki 1978,1989; May et al. 1975; Peletier 1978; Nagylaki and Moody 1980; Fife and Peletier 1981; Keller 1984).

Interestingly, these theoretical models have remained largely restricted to make predictions on the effect of different evolutionary forces on gene frequencies and very few attempts have been made at fitting them to data. A first attempt was made by Novembre et al. (2005) who used Fisher's (1937) wave of advance model as the basis for a maximum likelihood method to make inferences about the evolutionary history of the CCR5 $\Delta 32$ mutation. However, this approach remains largely ad-hoc. Given the many examples of genetic clines, it is desirable to develop a general estimation method that could be applied to any species. Here we use a theoretical model based on a partial diffusion equation as the basis for a hierarchical Bayesian formulation that allows us to estimate selection coefficients along an environmental gradient and other parameter of interest such as the heterozygous effect. The method is implemented using the Approximate Bayesian Computation framework and we investigate its performance with a simulation study. Finally, we provide an application example using the well studied case of the mummichog (*Fundulus heteroclitus*) cline at the LDH-B locus.

10.2 Models and Methods

In this section we describe models and numerical methods for the generation of synthetic genetic clines and their use within an ABC framework.

10.2.1 Migration-selection model for genetic clines

We consider a diploid species that is continuously distributed along a linear habitat Ω which exhibits an environmental gradient. We assume discrete and non-overlapping generations and random mating between individuals within each local population. We neglect mutation and random genetic drift and consider only migration and natural selection as evolutionary forces. The dispersal process is homogeneous in the habitat and independent of individuals while selection depends on the position in the habitat because of local environmental conditions. Under these assumptions we focus on the evolution of the spatial distribution of allele frequencies at a single diallelic locus (A: p, a: $q = 1 - p$).

In our model a single generation follows the four step life cycle represented in Figure 10.1: at each location individuals mate randomly and give birth to zygotes that undergo natural selection independently. Then the resulting offspring disperse and individuals contributing to reproduction start a new generation. Note that in our scheme all fertilities are the same and that selection occurs through viability.

We introduce selection through the relative fitness of each possible genotype

$$\begin{aligned} AA &: w_{AA} = 1 + s(x) \\ Aa &: w_{Aa} = 1 + hs(x) \\ aa &: w_{aa} = 1 \end{aligned} \tag{10.1}$$

where $s \in (-1, 1)$ is the selection coefficient and $h \in (0, 1)$ the level of dominance, i.e. the heterozygous effect. Because of habitat heterogeneities due to changing environmental conditions, the selection coefficient s depends on the position x in the habitat so that $s(x) = s_{\max}\gamma(x)$ with

$$\gamma(x) = \frac{2}{1 + \exp(-\underbrace{(a_0 + a_1 G^1(x) + \ldots + a_R G^R(x) + \xi)}_{=\alpha(x)})} - 1 \tag{10.2}$$

The regression model links selection with normalized environmental factor r at each location x, $G^r(x)$, weighted by its influence, a_r. The value for $s_{\max} \in (0, 1)$ measures the strength of selection and γ its spatial variation. In order to allow deviation from the exact regression we introduce Gaussian noise as $\xi \backsim \mathcal{N}(0, 1/\tau)$ where $\tau \backsim \chi^2$.

The rationale for the choice of $\gamma(x)$ (cf. eqn. (10.2)) is that we want the selection coefficient to ascertain $|s| < s_{\max} \leq 1$. Note that when there is a strong influence of local environmental factors, i.e. $\alpha(x) \to \pm\infty$, the strength of selection is close to its extremal value $s_{\max}$ because $\gamma(x) \to \pm 1$. On the other hand, when local constraints are weak, i.e. $\alpha(x) \to 0$, the selection coefficient is

weak and approximatively linearly dependent on the environmental variables as $\gamma(x) \approx \frac{\alpha(x)}{2} \to 0$.

Migration is parameterized by the diffusion coefficient $D = \sigma^2/2$, where σ^2 is the variance of the parent-offspring dispersal distribution. We assume that individuals can not disperse outside the habitat and that D is bounded and lies in $(0, D_{\max})$. Note that the value for $D_{\max}$ depends on the dispersal kernel and the habitat of the species under study but we assume that it is independent of the position.

10.2.2 Diffusion approximation for selection and migration

In the diffusion approximation the spatial (geographical position x) and temporal (time t) evolution of gene frequency is described by the partial differential equation (PDE) (Wright 1931, Fisher 1937, Haldane 1948, Nagylaki 1989)

$$\frac{\partial p}{\partial t} = \underbrace{D\frac{\partial^2 p}{\partial x^2}}_{\mathcal{L}(p)} + \underbrace{s(x)pq\frac{(1-h)p+hq}{(1+s)p^2+2(1+hs)pq+q^2}}_{\mathcal{G}(p)} \text{ in } \Omega \times (0,+\infty) \tag{10.3}$$

with the absence of gene flow at the limits of the habitat (homogeneous Neumann boundary conditions), i.e.

$$\frac{\partial p}{\partial x} = 0 \text{ on } \partial\Omega \times (0,+\infty) \tag{10.4}$$

and specified initial conditions $p(x,0) = p_0(x)$ in Ω. The length of the habitat is denoted by L and $T_{\max}$ is the upper bound for time. The deterministic model of equation (10.3) consists of a linear diffusion term, $\mathcal{L}$, and non-linear local growth term, $\mathcal{G}$. The former models the spatial variation of allele frequency due to migration and the later the temporal changes in allele frequency due to selection and local environmental factors.

The model for deterministic clines described by equations (10.2), (10.3) and (10.4) is parametrized by the parameter vector $\boldsymbol{\phi} = (D, h, \tau, (a_r))$. These equations can be solved by a numerical scheme and, thus, allow for the generation of synthetic allele frequencies for a given parameter set. Now that we have a model to generate spatial distributions of gene frequency, we explain how it can be used within an ABC approach to estimate dispersal and selection parameters.

10.2.3 Bayesian formulation

In the model we describe above parameters of interest are the diffusion coefficient D, the heterozygous effect h, the noise in the environmental gradient τ and

the influence of environmental factors on selection, a's. Note that the local selection coefficient and its variation are derived parameters that are also of interest. The hierarchical Bayesian model used in this study is represented by the directed acyclic graph (DAG) of Figure 10.3.

Given prior distributions for model parameters and the likelihood of the data, $\Pr(\boldsymbol{p}|D, h, \tau, \boldsymbol{a})$, Bayes'rule gives the posterior distribution of the parameters up to a multiplicative constant, i.e.

$$\Pr(D, h, \tau, \boldsymbol{a}|\boldsymbol{p}) \propto \Pr(\boldsymbol{p}|D, h, \tau, \boldsymbol{a}) \Pr(D) \Pr(h) \Pr(\tau) \Pr(\boldsymbol{a}) \tag{10.5}$$

where $\Pr(D)$, $\Pr(h)$, $\Pr(\tau)$ and $\Pr(\boldsymbol{a})$ are prior distributions for model parameters.

We assume a uniform prior on the interval $(0, D_{\max})$ for the dispersal parameter D. The level of dominance h follows a uniform prior on the unit interval. The prior for the influence of environmental factors, the a's, is a standard normal distribution. Finally we use an inverse chi-square prior distribution for the noise in the environmental gradient τ.

When the likelihood function is known, one can sample from the posterior distribution and get posterior estimates by using MCMC algorithms. Although possible in our case, the derivation of a likelihood function is difficult and would not rely on a biological model. The use of an ABC approach provides an attractive alternative as it avoids the use of a likelihood function.

Thus, following the DAG and using the numerical methods described in Appendix I, we can generate synthetic gene frequencies from various dispersal and selection parameters drawn from their prior distributions. The details of our ABC approach are given in the following section.

10.2.4 Approximate Bayesian Computation approach

ABC methods rely on the generation of synthetic data to estimate parameters of interest without the use of a likelihood function. Basically these methods consist of a rejection sampler: only synthetic data that are 'close' to observed data are used for estimation purposes. The term 'close' means that the euclidian distance between summary statistics computed for synthetic and observed data lies within a given tolerance limit δ.

Genetic clines become stable when the homogenizing effects of migration counterbalance the diversifying effects of natural selection. This equilibrium gene frequencies are simulated by our numerical solver by waiting for the numerical solution to reach stationarity, i.e. $\frac{\partial p}{\partial t} = 0$. Thus, the allele frequencies we simulate in our ABC framework, $\boldsymbol{\phi_j}$'s, from a given parameter set, $\boldsymbol{\phi}_j$'s, consist of the stationary solution of eqn. (10.3) at sampled locations.

Depending on the species under study, genetic clines observed in natural populations may be less smooth than the ones we can generate by solving eqn. (10.3).

More precisely, real data are the outcome of complex stochastic processes whereas our model is deterministic and does not take into account random effects (e.g. genetic drift) and assume constant effects of both migration and selection. For these reasons, we use smoothing splines (Green and Silverman 1994) of allele frequencies as summary statistics instead of using them directly. Note that smoothing splines depend on a smoothness parameter that controls a trade-off between data fitting and smoothing. The choice of this smoothness parameter will be discussed later in the text and in Appendix III.

We generate n independent pairs $(\boldsymbol{\phi}_j, \boldsymbol{p}_{(j)})$ where each $\boldsymbol{\phi_j}$ is an independent draw from the prior and the $\boldsymbol{p}_{(j)}$s are simulated values of $\boldsymbol{p}$ with $\boldsymbol{\phi} = \boldsymbol{\phi}_j$. The idea is that $\boldsymbol{\phi}_j$s for which $\|\boldsymbol{p}_{(j)} - \boldsymbol{p}\| \leq \delta$ provide an approximate random sample necessary to carry out posterior analysis.

Following Beaumont *et al* (2002)'s approach, statistical inferences are based on a local-linear regression of simulated parameter values on simulated summary statistics. Then the observed summary statistics are substituted into the regression equation and provide a sample from the approximate posterior distribution. Each simulated pair is weighted using a kernel function K_δ (e.g. Epanechnikov kernel) that decreases to zero when the euclidean distance between observed and simulated summary statistics increases.

The approximate posterior distribution of each parameter is computed from the weighted and adjusted sample by using standard kernel density estimation. Doing so we can calculate posterior mode and mean and 95% HPDIs as estimators of dispersal and selection parameters.

The two following sections present the results of our sensitivity analysis and, then, we demonstrate our method with Powers et al. (1978)'s *F. heteroclitus* data set.

10.3 Sensitivity analysis

We evaluated the sensitivity of the model by generating synthetic data using a discrete-time simulation model. Following the detailed life cycle of Figure 10.2, let p_i^k denote the frequency of allele A at location i and generation k. During reproduction random mating between breeders restore Hardy-Weinberg proportions but does not change allele frequencies which remain the same in the newly formed zygotes. Then, before migration, natural selection acts on viability at each location; the corresponding change in allele frequencies is

$$\delta p = pq\frac{(w_{AA} - w_{Aa})p + (w_{Aa} - w_{aa})q}{\bar{w}} = spq\frac{(1-h)p + hq}{(1+s)p^2 + 2(1+hs)pq + q^2} \quad (10.6)$$

where $\bar{w} = w_{AA}p^2 + 2w_{Aa}pq + w_{aa}q^2$ is the mean fitness. Then adjacent colonies exchange adults independently of their genotypes at rate $m/2$; the corresponding change in allele frequencies is

$$\delta p_i = \frac{m}{2}(p_{i-1} - 2p_i + p_{i+1}) \tag{10.7}$$

Finally each genotype contributes to the breeders' pool (we assume no selection on fecundity).

We considered equally spaced colonies distributed along a linear habitat that approximates the range of *F. heteroclitus* (cf. Figure 10.4). We assumed that the spatial distribution of selection coefficients was a linear function of the position in the habitat. The iterative process that simulates discrete clines was run for 1000 generations to ensure equilibrium between selection and migration. We assumed that the upper bound for the diffusion coefficient was similar to the one estimated for *F. heteroclitus* by Brown and Chapman (1991), i.e. $D_{\max} = 4$.

Each individual simulation of our ABC method generates a cline at migration-selection equilibrium. The stationarity is ascertained by setting $T_{\max} = 1000$ time units and checking that the two last outcomes of our numerical solver are close to each other. The final state is then approximated by a smoothing spline whose smoothness parameter is determined by cross-validation methods (cf. Appendix III). Note that because both the numerical solutions of eqn. (10.3) and the synthetic allele frequencies generated by the discrete model are smooth, the smoothing spline is almost the same as a data fit.

We chose a set of default values for the parameters of the discrete simulation model and then studied the effect of varying only one at time. These default values are $D = 2$ for the diffusion coefficient, $h = 0.5$ for the level of dominance, $\tau = 0$ for the noise in the environmental gradient, $a_0 = -0.1$ for the intercept of the regression and $a_1 = 0.9$ for the influence of the environmental gradient. We studied the effect of varying individually $D \in \{0.5, 1.0, 2.0, 3.0, 3.5\}$, $h \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, $\tau \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, $a_0 \in \{-0.1, -0.3, -0.5, -0.7, -0.9\}$, and $a_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ while keeping the other parameters set to their default values.

We also investigated the effect of varying the choice of the tolerance, δ, for the local-linear regression of our ABC scheme. The choice of δ involves a bias-variance trade-off as increasing the acceptance rate decreases variance but increases bias. In our method we set δ to be a quantile, q_δ, of the empirical distribution of distances between observed and simulated data. Thus, for the regression method, we assigned a nonzero weight to a proportion $q_\delta \in \{0.00050, 0.00075, 0.00100, 0.00125, 0.00150\}$ of the simulations that are closest to observed summary statistics.

We used our ABC scheme to estimate the parameters. We drew $n = 100000$ samples from the prior distributions and solved the PDE eq. (10.3). We then

accepted a fraction q_δ of the simulations that were closest to the observed values. Finally we applied the weighting and adjusting procedure of Beaumont et al. (2002). We were interested in the ability of our approach to provide accurate estimates of both dispersal and selection parameters. We consider both mode and mean of the approximate posterior densities as point estimates. The RMSEs of both estimators were computed by running 50 replicates for each synthetic data set to evaluate the accuracy of our method (bias-variance trade-off). We also calculated the 95%HPDI of the approximate posterior distributions of each parameter as interval estimates.

10.3.1 Results

We investigate the effect of varying each parameter individually on the approximate posterior estimates of dispersal and selection parameters.

The effect of varying tolerance, q_δ, on the estimates is presented in Table 10.1. Varying q_δ has a slight effect on the bias of estimates. Increasing the number of accepted simulations first increases and then decreases the accuracy of D, h and a_1 estimates as shown by their RMSEs. Conversely, the RMSEs for the deviation τ and the intercept a_0 of the regression for the selection coefficients first increase and then decrease when increasing the acceptance rate. Note that the RMSE values for τ are very high (because the true value is very small) whereas other parameter estimates are much more accurate. Increasing accepted simulations leads to wider HPDIs for all estimated parameters indicating a decrease in precision.

In order to investigate the sensitivity of the estimates to model parameters, we fixed the tolerance $q_\delta = 0.00075$. The effect of varying the diffusion coefficient, D, on posterior estimates is presented in Table 10.2. The mode of the posterior distribution of D slightly overestimates the diffusion coefficient when $D \neq 2$; the mean provides underestimates of the dispersal parameter when $D \geq 2$ and overestimates for smaller values. The posterior mode of D is more accurate than the mean when $D \leq 2$ as indicated by the RMSEs. The true values of the diffusion coefficient always fall within the HPDIs whose widths increase when increasing D. The heterozygous effect is underestimated when $D \leq 2$ and, as shown by RMSEs, the mean is more accurate than the mode also when $D \leq 2$. The true values of h always fall within the HPDIs whose widths decrease when increasing D. The biases of both τ estimates increase when increasing D but the mode performs better than the mean; the widths of the HPDIs increase and do not change that much when $D \geq 2$. Both mode and mean tend to underestimate the regression parameters; increasing D leads to narrower HPDIs for the posterior distributions of the a's which always contain the true values.

The effect of varying the heterozygous effect, h, on posterior estimates is presented in Table 10.3. The diffusion coefficient is slightly underestimated, the mode

being more accurate than the mean when $h \geq 0.25$. The HPDIs for D, whose widths are almost constant, always contain the true values of the diffusion coefficient. Both mean and mode underestimate the heterozygous effect when $h < 0.90$, but the former estimate is more accurate than the latter. The HPDIs for h, whose widths are almost constant, always contain the true value. We observe the same pattern as when varying D for the noise in the environmental gradient (cf. previous paragraph). The intercept of the regression is underestimated when $h \geq 0.25$ while the influence of the environmental gradient is underestimated when $h \leq 0.75$. For both regression parameters, the mean and the mode show almost the same accuracy, and their true values always fall within the HPDIs.

The effect of varying the noise in the environmental gradient, τ, on posterior estimates is presented in Table 10.4. The diffusion coefficient is underestimated when $\tau < 10^{-2}$ and for larger values. This is the case for both mode and mean but the former is in general more accurate than the latter. Although the HPDIs always contain the true values of D, the RMSE increases steadily as τ increases; the same pattern is observed for the heterozygous effect. The effect of increasing the noise in the regression does not have much an effect on τ point estimates, which are accurate only when $\tau = 10^{-3}$. The RMSEs decrease sharply (2 orders of magnitude for $\tau \leq 10^{-3}$) and then increase slowly while increasing the noise in the environmental gradient; the RMSEs of the mode of τ are lower than those of the mean indicating that the mode performs better than the mean. The true values of τ fall within the HPDIs only when $\tau \geq 10^{-3}$ and their widths increase when increasing the noise in the regression. The intercept of the regression is overestimated only when $\tau = 10^{-2}$ and is underestimated otherwise; both point estimates of a_0 have the same accuracy. The influence of the environmental factor is underestimated when $\tau \leq 10^{-2}$ and becomes overestimated otherwise. The mode of a_1 is slightly less accurate than the mean when $\tau < 10^{-2}$. The RMSEs of regression coefficients decrease while increasing the noise in the regression indicating that the estimates become less accurate. The HPDIs for the regression coefficients always contain the true value of these parameters and their widths are almost constant.

The effect of varying the regression coefficients, a_0 and a_1, for the selection parameter on posterior estimates is presented in Tables 10.5 and 10.6. Varying the intercept (Table 10.5), a_0 , does not have a clear effect on the mean, mode and RMSE of the diffusion coefficient as long as it remains larger than -0.9. For smaller values, the accuracy and precision degrades pronouncedly and the true value is not included in the HPDIs. On the other hand, the estimates of h remain fairly accurate and precise for all values of the intercept. Estimates of the noise in the environmental gradient τ do not exhibit a clear pattern as a_0 decreases. The estimates of a_0 are always fairly accurate and precise while those of a_1 tend to be underestimates when $a_0 \leq -0.3$ and slight overestimates otherwise. Note,

however, that the HPDIs always contain the true value. Varying the slope a_1 of the regression, that is the steepness of the selection gradient has important effects on the estimates of the diffusion coefficient (Table 10.6). When a_1 is small ($= 0.1$), all parameters are strongly underestimated but the accuracy and precision increases rapidly for larger values and the true values are always included in the HPDIs except for the estimate of τ, that is always poorly estimated.

Overall, these results indicate that with the exception of the environmental noise, all parameters can be fairly well estimated under most circumstances. This is particularly the case for the three parameters that are most interesting, namely the diffusion coefficient, the heterozygous effect and the slope of the selection gradient.

10.4 Application to *Fundulus heteroclitus* data

In order to exemplify the application of the method, we analyse data on the well known LDH-B cline observed in the mummichog, *Fundulus heteroclitus* (Powers and Place 1978). This species inhabits salt marshes along the Atlantic coast of North America with a range that extends from northern Florida to the Gulf of St Lawrence. It is considered as a model for the study of intraspecific geographical variation (e.g. Avise 2004) because of extensive latitudinal variation in a number of morphological, physiological and biochemical traits. Many allozyme loci exhibit a latitudinal cline in gene frequencies (Powers and Place 1978; Cashon et al. 1981; Powers et al 1986; Ropson et al. 1990). These clinal patterns are highly concordant and centered at 40.8°N latitude, roughly corresponding to northern New Jeresey, where the Hudson river enters the Atlantic Ocean. Additionally, these clines are associated with a directional change in mean water temperature, which is highly correlated ($r = 0.99$) with latitude, demonstrating a 1°C change per degree latitude (Powers and Place, 1978).

Biochemical data on the LDH-B allozymes indicate a possible selective advantage for the Ldh-B^hB^h allotype in cold water and for the Ldh-B^aB^a in warm habitats (Powers and Place 1978). We therefore use our method to estimate the selective gradient, the heterozygous effect and the variance of the parent-offspring dispersal distribution.

The *F. heteroclitus* habitat is assimilated to the best-fitting great circle of Powers and Place (1978) and Adams et al. (2006)' sampled locations (see Figure 10.4). The best-fit great circle and the projections of sampling sites onto this circle are computed using Wessel and Smith (1998)'s generic mapping tools.

The ABC simulations were run using the same settings as for the sensitivity analysis: we generated $n = 100000$ clines with the same level of tolerance q_δ for estimation purposes. We investigated the effect of varying the smoothness

parameter for smoothing splines on posterior estimates of dispersal and selection parameter. The value for the upper bound of diffusion coefficient $D_{\max} = 4$ was chosen according to Brown and Chapman (1991)'s analysis of mummichog mtDNA data in the absence of strong selection. The value for $s_{\max}$ was tuned to avoid the simulation of steep clines that would have been rejected and, thus, useless for estimation purposes. Our pilot study led us to choose $s_{\max} = 0.01$ to simulate clines whose shapes resemble the observed cline.

The results for the analyses of *F. heteroclitus* allozyme data are shown in Tables 10.7 and 10.8. When using observed allele frequencies (Table 10.7), estimates of the diffusion coefficient and the level of dominance depend strongly on the tolerance level; the effect on regression parameters τ, a_0 and a_1 is less pronounced. The same pattern is observed when using smoothing splines of allele frequencies (Table 10.8), but the effect is somewhat weaker. Additionally, there are important differences between the estimates obtained using observed allele frequencies and those obtained using smoothing splines. Although in the case of the regression parameters and, to a lesser extent the diffusion coefficient, there is extensive overlap between the HPDIs obtained using the two approaches, point estimates are fairly different. Given these differences it seems more appropriate to use the results obtained from the raw allele frequencies to draw inferences. Additionally, based on the simulation study that showed a minimum in RMSE for $q_\delta = 0.00075$ we chose this quantile's results for inference purposes. With these caveats in mind it is possible to conclude that there is a strong environmental gradient ($a_1 = -0.225$) and that the dispersal dispersal distances are smaller than those reported by Brown and Chapman (1991). Finally, the heterozygous effect is additive ($h \in (0.15; 0.61)$).

10.5 Discussion

We present an ABC method for the estimation of dispersal and selection parameters based on a partial differential equation that models genetic clines under migration and selection. Our approach estimates the variance of the parent-offspring dispersal distribution, the heterozygous effect and the influence of environmental factors on the selection coefficients. Observed data consist of allele frequencies at a biallelic locus and values for environmental factors across the studied species range.

We generated synthetic data following a discrete generation model to study the performance of our method under different scenarios. Although limited, our simulation study indicates that accurate estimates can be obtained when the environmental gradient is well established. Future simulation studies should consider more complex scenarios with several environmental factors.

We demonstrate our method through the analysis of *Fundulus heteroclitus* al-

lozyme data considering latitude as an explanatory factor for the variation of selection coefficients at the LDH-B locus. This provided us with the opportunity of highlighting some of the problems that arise when trying to apply the method to real data. More specifically, the output of the simulation step in our ABC approach is a smooth function whereas observed data can be rather noisy. A practical answer to noise elimination in experimental data is to use splines for smoothing purposes. This approach is very popular in disciplines such as image processing and statistics. In biology it has been applied by Ballain et al. (1998) to fluorescence signals for the analysis of cortex activity in rats. Following the example of these studies, we used smoothing splines to filter the noise present in observed allele frequencies. The main challenge with smoothing splines is to set an appropriate value for their smoothing parameters. A smoothing parameter that is 'too low' will yield a fit practically equivalent to the data, and, conversely, a smoothness parameter that is 'too high' will produce a fit practically equivalent to the linear regression estimate of the data. In this study we use the leave-one-out approach that belongs to cross validation (CV) and generalized cross-validation (GCV) (Craven and Wahba 1979). An alternative method is that of Wahba (1985), which consists of a Generalized Maximum Likelihood (GLM) criterion and that could also be used to achieve spline smoothing of allele frequencies. It should be noted, however, that it is always important to explore the effect of using smoothing splines on the estimates. Ideally, they should allow us to reduce the variance but should not have a large effect on the posterior distribution. If the impact is too large, for example by changing the support of the posterior, then it is advisable to use the raw instead of the smoothed data. This is what we did in the case of *F. heteroclitus*.

The choice of $D_{\max}$ and $s_{\max}$ is very important because these two values control the shape of simulated clines in our ABC framework. They must be pilot tuned to avoid a high rejection rates in the ABC method. Note that the upper bound for the diffusion coefficient, $D_{\max}$, can be roughly approximated based on knowledge of the biology of the species under study. The choice of $s_{\max}$ is harder and requires a pilot study to avoid the simulation of step clines for which the effects of selection strongly overcome those of migration.

The biological interpretation for the intercept of the regression on selection coefficients is rather difficult. It can be though as an averaged level of selection that indicates if the considered allele is more or less advantageous than the other across the studied area. Alternatively, it can be considered as the selection expected if there was no environmental gradient.

Our study provides an extension for Novembre et al. (2005)'s approach for the estimation of dispersal and selection parameter. They considered an additive model for selection and allowed for the effect of environmental gradient on selection; here we also estimate the heterozygous effect by introducing the level of dominance

h. Additionnally, Novembre et al. (2005) estimates are based on a maximum likelihood method through a binomial sampler whereas we used our numerical solver in the likelihood free framework of ABC which avoid the derivation of a likelihood function.

Future developments of our method would improve the introduction of environmental factors as an hyperprior for selection coefficients (and not directly as a prior). Indeed, our experience in such problems lead us to think that it would reduce inference problems when estimating regression parameters, especially the noise in the regression. The method we introduce can also be extended to two-dimensional habitats, which is the case in most species. Another important aspect to study is the influence of varying the smoothing parameter or the use of other noise elimination methods since we showed that this does affect the estimates.

In several fields of biology many phenomena are modelled with stochastic or deterministic differential equations that depends on parameters of interest. Although population geneticists have developed such predictive tools, few studies use them for estimation purposes. However, in related fields of kinetics and ecology, Bayesian approaches have been successfully used for the estimation of kinetic constant rates (Golightly and Wilkinson 2006) and the rate of spread of a species (Wikle and Hooten 2005) by fitting models to observed data. More recently, Ramsay et al. (2007) proposed another promising approach for the estimation of differential equation parameters. With such powerful statistical tools, population genetics will be able to use existing theory to infer evolutionary parameters.

10.6 Tables

Effect of varying the tolerance q_δ on dispersal and selection parameter estimates.

Parameter	D	h	τ	a_0	a_1
True value	2.000	0.5000	10^{-5}	-0.1000	0.9000
$q_\delta = 0.00050$	1.8766 (0.0039)[1]	0.4968 (0.0001)	0.0011 ($1.1228e+4$)	-0.1045 (0.0023)	0.8916 (0.0001)
	1.8783 (0.0038)[2]	0.4980 (0.0000)	0.0012 ($1.3061e+4$)	-0.1026 (0.0010)	0.8900 (0.0001)
	[1.6328; 2.1189][3]	[0.4697; 0.5276]	[0.0005; 0.0017]	[-0.1186; -0.0817]	[0.8338; 0.9401]
$q_\delta = 0.00075$	1.9185 (0.0018)	0.4963 (0.0001)	0.0011 ($1.2154e+4$)	-0.1023 (0.0007)	0.8624 (0.0018)
	1.8836 (0.0035)	0.4969 (0.0001)	0.0012 ($1.5356e+4$)	-0.1017 (0.0005)	0.8695 (0.0012)
	[1.6019; 2.1579]	[0.4630; 0.5339]	[0.0005; 0.0021]	[-0.1231; -0.0803]	[0.7858; 0.9557]
$q_\delta = 0.00100$	1.9004 (0.0026)	0.4944 (0.0001)	0.0011 ($1.2212e+4$)	-0.1015 (0.0005)	0.8424 (0.0041)
	1.8807 (0.0036)	0.4957 (0.0001)	0.0012 ($1.5139e+4$)	-0.1010 (0.0003)	0.8599 (0.0020)
	[1.5483; 2.1925]	[0.4556; 0.5418]	[0.0004; 0.0021]	[-0.1242; -0.0739]	[0.7643; 0.9745]
$q_\delta = 0.00125$	1.8595 (0.0051)	0.4905 (0.0004)	0.0011 ($1.0972e+4$)	-0.1004 (0.0002)	0.8321 (0.0058)
	1.8629 (0.0048)	0.4924 (0.0002)	0.0012 ($1.4847e+4$)	-0.1001 (0.0002)	0.8537 (0.0027)
	[1.4801; 2.2574]	[0.4480; 0.5393]	[0.0004; 0.0022]	[-0.1268; -0.0677]	[0.7407; 0.9975]
$q_\delta = 0.00150$	1.7870 (0.0115)	0.4848 (0.0009)	0.0010 ($0.9316e+4$)	-0.1025 (0.0009)	0.8277 (0.0065)
	1.8236 (0.0079)	0.4892 (0.0005)	0.0012 ($1.4227e+4$)	-0.0994 (0.0003)	0.8484 (0.0034)
	[1.3678; 2.3010]	[0.4407; 0.5433]	[0.0004; 0.0023]	[-0.1292; -0.0619]	[0.6977; 1.0439]

Table 10.1

a. mode estimate (RMSE)
b. mean estimate (RMSE)
c. [HPDI]

Effect of varying the diffusion coefficient D.

Parameter	D	h	τ	a_0	a_1
True value		0.5000	0.0000	−0.1000	0.9000
$D = 0.5000$	0.5091 (0.0003)[1]	0.4844 (0.0010)	0.0005	−0.0973 (0.0007)	0.8706 (0.0011)
	0.5086 (0.0003)[2]	0.4866 (0.0007)	0.0006	−0.0916 (0.0071)	0.8592 (0.0021)
	[0.4662; 0.5504][3]	[0.4257; 0.5432]	[0.0002; 0.0012]	[−0.1285; −0.0491]	[0.6776; 1.0153]
$D = 1.0000$	1.0068 (0.0000)	0.4952 (0.0001)	0.0007	−0.0893 (0.0115)	0.7757 (0.0191)
	1.0103 (0.0001)	0.4945 (0.0001)	0.0009	−0.0925 (0.0056)	0.7996 (0.0124)
	[0.9376; 1.0860]	[0.4482; 0.5397]	[0.0003; 0.0017]	[−0.1251; −0.0611]	[0.6684; 0.9290]
$D = 2.0000$	1.9219 (0.0015)	0.4965 (0.0000)	0.0011	−0.1014 (0.0002)	0.8647 (0.0015)
	1.8847 (0.0033)	0.4971 (0.0000)	0.0013	−0.1017 (0.0003)	0.8694 (0.0012)
	[1.6024; 2.1585]	[0.4634; 0.5338]	[0.0005; 0.0021]	[−0.1226; −0.0802]	[0.7895; 0.9550]
$D = 3.0000$	3.1258 (0.0018)	0.5115 (0.0005)	0.0011	−0.0902 (0.0097)	0.8734 (0.0009)
	2.9880 (0.0000)	0.5118 (0.0006)	0.0012	−0.0896 (0.0109)	0.8731 (0.0009)
	[2.3235; 3.9525]	[0.4746; 0.5497]	[0.0004; 0.0020]	[−0.1142; −0.0620]	[0.8055; 0.9348]
$D = 3.5000$	3.6268 (0.0013)	0.5081 (0.0003)	0.0010	−0.0905 (0.0091)	0.8834 (0.0003)
	3.4798 (0.0000)	0.5101 (0.0004)	0.0011	−0.0917 (0.0068)	0.8806 (0.0005)
	[2.9431; 3.9687]	[0.4739; 0.5491]	[0.0005; 0.0019]	[−0.1151; −0.0659]	[0.8137; 0.9435]

Table 10.2

a. mode estimate (RMSE)
b. mean estimate (RMSE)
c. [HPDI]

Effect of varying the level of dominance h.

Parameter	D	h	τ	a_0	a_1
True value	2.0000		0.0000	−0.1000	0.9000
$h = 0.1000$	1.8188 (0.0082)[1]	0.0802 (0.0394)	0.0007	−0.0872 (0.0165)	0.7933 (0.0141)
	1.8405 (0.0064)[2]	0.0835 (0.0271)	0.0008	−0.0871 (0.0166)	0.7967 (0.0132)
	[1.5218; 2.1315][3]	[0.0145; 0.1263]	[0.0003; 0.0012]	[−0.1228; −0.0572]	[0.6169; 0.9697]
$h = 0.2500$	1.8795 (0.0036)	0.2478 (0.0001)	0.0009	−0.1029 (0.0008)	0.8675 (0.0013)
	1.8720 (0.0041)	0.2520 (0.0001)	0.0009	−0.1034 (0.0012)	0.8659 (0.0014)
	[1.5991; 2.1786]	[0.2113; 0.3012]	[0.0003; 0.0015]	[−0.1496; −0.0649]	[0.6744; 1.0317]
$h = 0.5000$	1.9219 (0.0015)	0.4965 (0.0000)	0.0011	−0.1014 (0.0002)	0.8647 (0.0015)
	1.8847 (0.0033)	0.4971 (0.0000)	0.0013	−0.1017 (0.0003)	0.8694 (0.0012)
	[1.6024; 2.1585]	[0.4634; 0.5338]	[0.0005; 0.0021]	[−0.1226; −0.0802]	[0.7895; 0.9550]
$h = 0.7500$	1.8860 (0.0032)	0.7408 (0.0002)	0.0009	−0.0981 (0.0003)	0.8660 (0.0014)
	1.8739 (0.0040)	0.7433 (0.0001)	0.0011	−0.0990 (0.0001)	0.8648 (0.0015)
	[1.6192; 2.1528]	[0.7157; 0.7756]	[0.0004; 0.0020]	[−0.1192; −0.0793]	[0.7670; 0.9650]
$h = 0.9000$	1.8665 (0.0045)	0.9209 (0.0005)	0.0010	−0.0982 (0.0003)	0.9160 (0.0003)
	1.8463 (0.0059)	0.9185 (0.0004)	0.0014	−0.0977 (0.0005)	0.9194 (0.0005)
	[1.4542; 2.0816]	[0.8814; 0.9896]	[0.0004; 0.0024]	[−0.1156; −0.0770]	[0.7553; 1.0927]

Table 10.3

a. mode estimate (RMSE)
b. mean estimate (RMSE)
c. [HPDI]

Effect of varying the noise in the environmental gradient τ.

Parameter	D	h	τ	a_0	a_1
True value	2.0000	0.5000		-0.1000	0.9000
$\tau = 0$	1.9219 (0.0015)[1]	0.4965 (0.0000)	0.0011	-0.1014 (0.0002)	0.8647 (0.0015)
	1.8847 (0.0033)[2]	0.4971 (0.0000)	0.0013	-0.1017 (0.0003)	0.8694 (0.0012)
	[1.6024; 2.1585][3]	[0.4634; 0.5338]	[0.0005; 0.0021]	[-0.1226; -0.0802]	[0.7895; 0.9550]
$\tau = 10^{-5}$	1.9185 (0.0018)	0.4963 (0.0001)	0.0011 ($1.2154e+4$)	-0.1023 (0.0007)	0.8624 (0.0018)
	1.8836 (0.0035)	0.4969 (0.0001)	0.0012 ($1.5356e+4$)	-0.1017 (0.0005)	0.8695 (0.0012)
	[1.6019; 2.1579]	[0.4630; 0.5339]	[0.0005; 0.0021]	[-0.1231; -0.0803]	[0.7858; 0.9557]
$\tau = 10^{-4}$	1.9126 (0.0032)	0.4974 (0.0003)	0.0011 ($1.0410e+2$)	-0.1014 (0.0022)	0.8634 (0.0020)
	1.8891 (0.0043)	0.4979 (0.0002)	0.0013 ($1.3520e+2$)	-0.1010 (0.0021)	0.8701 (0.0015)
	[1.6112; 2.1590]	[0.4638; 0.5372]	[0.0005; 0.0021]	[-0.1224; -0.0790]	[0.7844; 0.9581]
$\tau = 10^{-3}$	1.9366 (0.0527)	0.4976 (0.0032)	0.0011 (0.2256)	-0.0995 (0.0426)	0.8620 (0.0087)
	1.9352 (0.0421)	0.4977 (0.0032)	0.0013 (0.3896)	-0.0992 (0.0440)	0.8682 (0.0081)
	[1.5307; 2.4013]	[0.4625; 0.5363]	[0.0005; 0.0022]	[-0.1259; -0.0736]	[0.7683; 0.9658]
$\tau = 10^{-2}$	2.2167 (0.4610)	0.5017 (0.1370)	0.0068 (4.1791)	-0.1186 (1.4245)	0.9476 (0.1194)
	2.2100 (0.3962)	0.5015 (0.1338)	0.0078 (5.2847)	-0.1183 (1.4293)	0.9521 (0.1209)
	[1.6884; 2.8265]	[0.4482; 0.5524]	[0.0031; 0.0128]	[-0.1457; -0.0906]	[0.8528; 1.0483]

Table 10.4

a. mode estimate (RMSE)
b. mean estimate (RMSE)
c. [HPDI]

Effect of varying the intercept of the regression for the selection coefficients a_0.

Parameter	D	h	τ	a_0	a_1
True value	2.0000	0.5000	0.0000		0.9000
$a_0 = -0.1000$	1.9219 (0.0015)[1]	0.4965 (0.0000)	0.0011	−0.1014 (0.0002)	0.8647 (0.0015)
	1.8847 (0.0033)[2]	0.4971 (0.0000)	0.0013	−0.1017 (0.0003)	0.8694 (0.0012)
	[1.6024; 2.1585][3]	[0.4634; 0.5338]	[0.0005; 0.0021]	[−0.1226; −0.0802]	[0.7895; 0.9550]
$a_0 = -0.3000$	1.8184 (0.0082)	0.4935 (0.0002)	0.0008	−0.2993 (0.0000)	0.8827 (0.0004)
	1.8225 (0.0079)	0.4966 (0.0000)	0.0009	−0.2975 (0.0001)	0.8860 (0.0002)
	[1.1210; 2.3045]	[0.4706; 0.5312]	[0.0005; 0.0014]	[−0.3262; −0.2580]	[0.8278; 0.9318]
$a_0 = -0.5000$	1.9952 (0.0000)	0.5101 (0.0004)	0.0009	−0.5023 (0.0000)	0.9019 (0.0000)
	1.9844 (0.0001)	0.5116 (0.0005)	0.0010	−0.4929 (0.0002)	0.9025 (0.0000)
	[1.3403; 2.4089]	[0.4826; 0.5434]	[0.0005; 0.0016]	[−0.5467; −0.4208]	[0.7970; 0.9957]
$a_0 = -0.7000$	1.9842 (0.0001)	0.5097 (0.0004)	0.0011	−0.7161 (0.0005)	0.9308 (0.0012)
	2.0273 (0.0002)	0.5103 (0.0004)	0.0011	−0.7225 (0.0010)	0.9265 (0.0009)
	[1.4951; 2.5956]	[0.4459; 0.5735]	[0.0004; 0.0019]	[−0.9111; −0.5212]	[0.6709; 1.1568]
$a_0 = -0.9000$	0.7662 (0.3806)	0.5020 (0.0000)	0.0014	−0.9089 (0.0001)	0.9145 (0.0003)
	1.3112 (0.1186)	0.5063 (0.0002)	0.0014	−0.9110 (0.0001)	0.9173 (0.0004)
	[0.2026; 3.9460]	[0.4601; 0.5606]	[0.0004; 0.0023]	[−0.9755; −0.8523]	[0.8679; 0.9674]

Table 10.5

a. mode estimate (RMSE)
b. mean estimate (RMSE)
c. [HPDI]

Effect of varying the influence of the environmental gradient a_1.

Parameter	D	h	τ	a_0	a_1
True value	2.0000	0.5000	0.0000	−0.1000	
$a_1 = 0.1000$	0.5808 (0.5035)[1]	0.2576 (0.2350)	0.0008	−0.0773 (0.0515)	0.0953 (0.0022)
	1.1403 (0.1848)[2]	0.2855 (0.1840)	0.0010	−0.0789 (0.0446)	0.0991 (0.0001)
	[0.0414; 2.8155][3]	[0.0322; 0.4998]	[0.0004; 0.0018]	[−0.1628; −0.0132]	[0.0350; 0.1724]
$a_1 = 0.3000$	1.8216 (0.0080)	0.5354 (0.0050)	0.0009	−0.0888 (0.0125)	0.2892 (0.0013)
	1.8454 (0.0060)	0.5361 (0.0052)	0.0010	−0.0893 (0.0114)	0.2923 (0.0007)
	[0.7282; 2.8809]	[0.4754; 0.5926]	[0.0005; 0.0017]	[−0.1309; −0.0620]	[0.2536; 0.3382]
$a_1 = 0.5000$	1.8913 (0.0030)	0.4997 (0.0000)	0.0010	−0.0992 (0.0001)	0.4858 (0.0008)
	1.9126 (0.0019)	0.4999 (0.0000)	0.0011	−0.0984 (0.0003)	0.4854 (0.0008)
	[1.5889; 2.2402]	[0.4590; 0.5405]	[0.0006; 0.0017]	[−0.1196; −0.0763]	[0.4361; 0.5405]
$a_1 = 0.7000$	1.8688 (0.0043)	0.4871 (0.0007)	0.0008	−0.1046 (0.0021)	0.6776 (0.0010)
	1.8849 (0.0033)	0.4891 (0.0005)	0.0009	−0.1022 (0.0005)	0.6849 (0.0005)
	[1.6109; 2.1911]	[0.4482; 0.5328]	[0.0004; 0.0015]	[−0.1319; −0.0738]	[0.6201; 0.7619]
$a_1 = 0.9000$	1.9219 (0.0015)	0.4965 (0.0000)	0.0011	−0.1014 (0.0002)	0.8647 (0.0015)
	1.8847 (0.0033)	0.4971 (0.0000)	0.0013	−0.1017 (0.0003)	0.8694 (0.0012)
	[1.6024; 2.1585]	[0.4634; 0.5338]	[0.0005; 0.0021]	[−0.1226; −0.0802]	[0.7895; 0.9550]

Table 10.6

a. mode estimate (RMSE)
b. mean estimate (RMSE)
c. [HPDI]

Parameter	D	h	τ	a_0	a_1
$q_\delta = 0.00050$	2.7840	0.0037	0.0002	0.0081	-0.1239
	2.7174	0.0047	0.0003	0.0082	-0.1298
	$[2.2705; 3.1508]$	$[0.0023; 0.0081]$	$[0.0001; 0.0004]$	$[-0.0126; 0.0234]$	$[-0.1994 - 0.0576]$
$q_\delta = 0.00075$	0.9299	0.3631	0.0006	0.0409	-0.2554
	1.0295	0.3863	0.0008	0.0418	-0.2614
	$[0.5712; 1.5449]$	$[0.1538; 0.6073]$	$[0.0003; 0.0015]$	$[0.0168; 0.0677]$	$[-0.3537 - 0.1784]$
$q_\delta = 0.00100$	0.5303	0.6024	0.0015	0.0377	-0.2721
	0.6001	0.5771	0.0022	0.0358	-0.2894
	$[0.2358; 0.9843]$	$[0.2978; 0.8982]$	$[0.0008; 0.0044]$	$[0.0095; 0.0647]$	$[-0.3996 - 0.1944]$
$q_\delta = 0.00125$	0.4711	0.5347	0.0017	0.0377	-0.2874
	0.5330	0.5128	0.0027	0.0382	-0.3099
	$[0.2011; 0.9100]$	$[0.2316; 0.8631]$	$[0.0010; 0.0053]$	$[0.0115; 0.0688]$	$[-0.4598 - 0.2113]$
$q_\delta = 0.00150$	0.3816	0.5030	0.0019	0.0445	-0.3094
	0.4311	0.4927	0.0032	0.0447	-0.3342
	$[0.1476; 0.7717]$	$[0.2160; 0.8576]$	$[0.0012; 0.0063]$	$[0.0162; 0.0777]$	$[-0.4782 - 0.2300]$

Table 10.7: Posterior estimates of dispersal and selection parameters for *F. heteroclitus* data when using observed allele frequencies.

a. mode estimate
b. mean estimate
c. [HPDI]

Parameter	D	h	τ	a_0	a_1
$q_\delta = 0.00050$	2.2502	0.0322	0.0039	0.0070	-0.2278
	2.1887	0.0439	0.0044	0.0074	-0.2327
	$[1.6346; 2.6698]$	$[0.0139; 0.0851]$	$[0.0020; 0.0072]$	$[-0.0132; 0.0336]$	$[-0.2991 - 0.1666]$
$q_\delta = 0.00075$	1.9429	0.0212	0.0016	0.0190	-0.1755
	1.8756	0.0348	0.0024	0.0168	-0.1899
	$[1.1899; 2.5744]$	$[0.0084; 0.0777]$	$[0.0009; 0.0045]$	$[-0.0106; 0.0461]$	$[-0.2979 - 0.1127]$
$q_\delta = 0.00100$	1.4543	0.1007	0.0020	0.0327	-0.2361
	1.4604	0.1376	0.0029	0.0317	-0.2585
	[0.76412.1955]	[0.03780.2893]	[0.00100.0056]	[0.00500.0603]	$[-0.3705 - 0.1708]$
$q_\delta = 0.00125$	1.4405	0.1117	0.0020	0.0339	-0.2530
	1.5068	0.1523	0.0031	0.0328	-0.2790
	[0.75992.2855]	[0.03490.2960]	[0.00110.0061]	[0.00710.0632]	$[-0.3955 - 0.1873]$
$q_\delta = 0.00150$	1.3443	0.1132	0.0020	0.0316	-0.2621
	1.4639	0.1562	0.0032	0.0329	-0.2860
	[0.69352.2602]	[0.01920.2941]	[0.00120.0061]	[0.00580.0652]	$[-0.4286 - 0.1899]$

Table 10.8: Posterior estimates of dispersal and selection parameters for *F. heteroclitus* data when using smoothing spline of observed allele frequencies.

a. mode estimate
b. mean estimate
c. [HPDI]

10.7 Figures

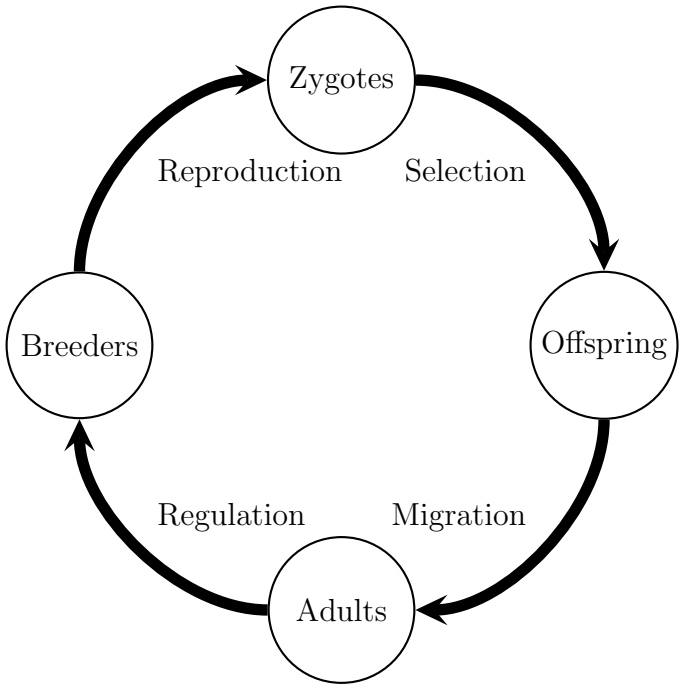


Figure 10.1: Diagram of the life cycle for the discrete model: the next generation is obtained from the previous one through the processes described in the text.

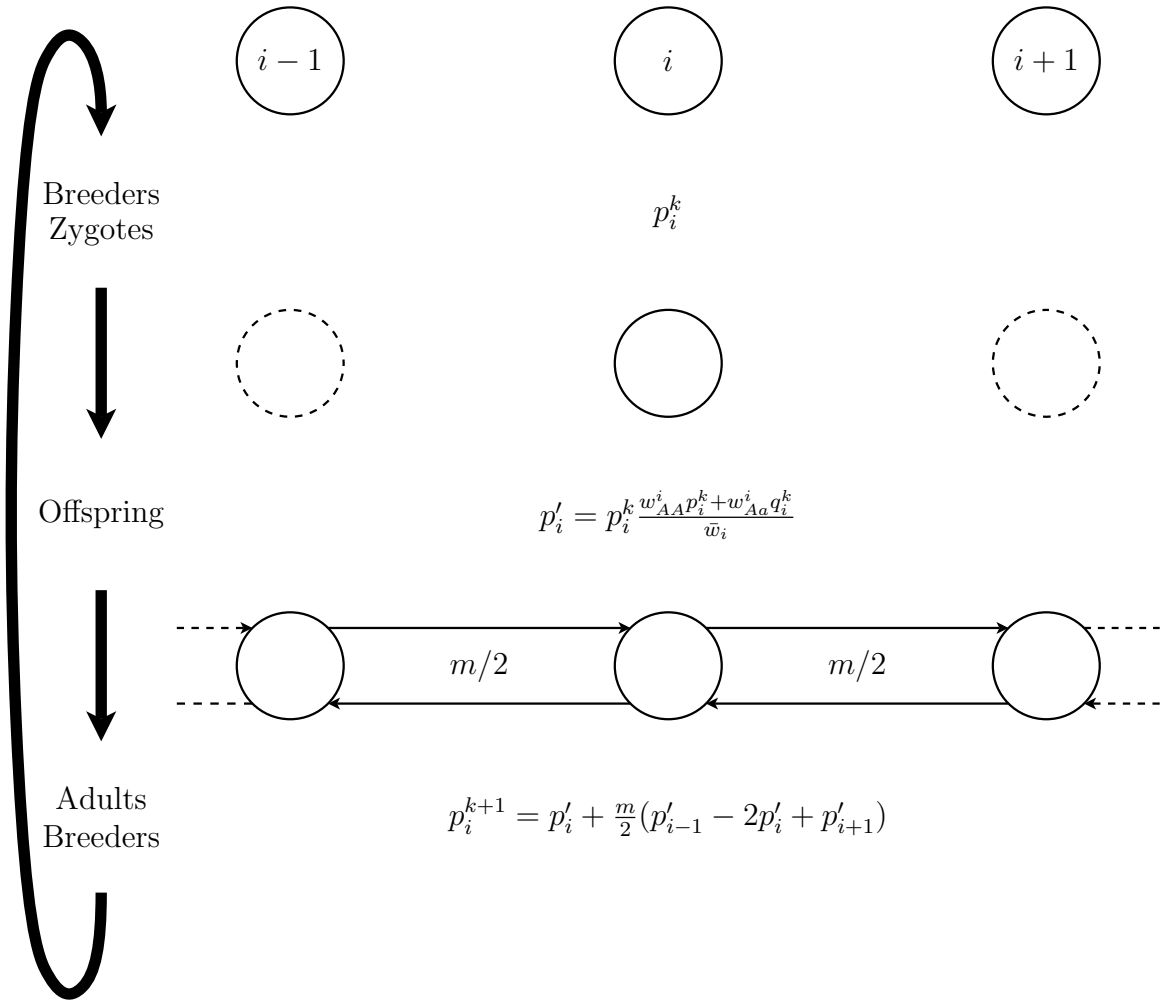


Figure 10.2: Details and assumptions for the discrete model. Breeders mate randomly and selection acts through viability of zygotes. Adjacent colonies exchange adults at rate $m/2$ (one dimensional stepping stone model). Changes in allele frequencies are given by equations (10.6) and (10.7).

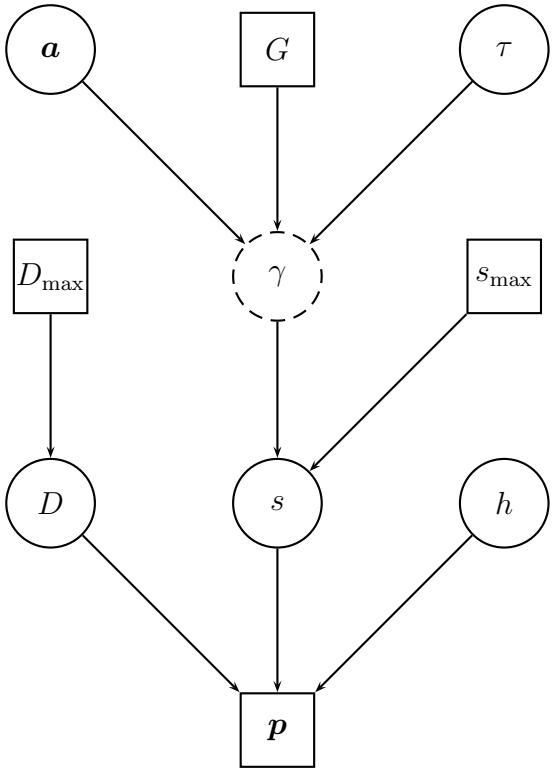


Figure 10.3: The DAG for the hierarchical Bayesian model. Square nodes denote known quantities (data) and circles represent parameters to be estimated. Lines between nodes represent direct stochastic relationships within the model. The variables within each node correspond to the different model parameters discussed in the text.

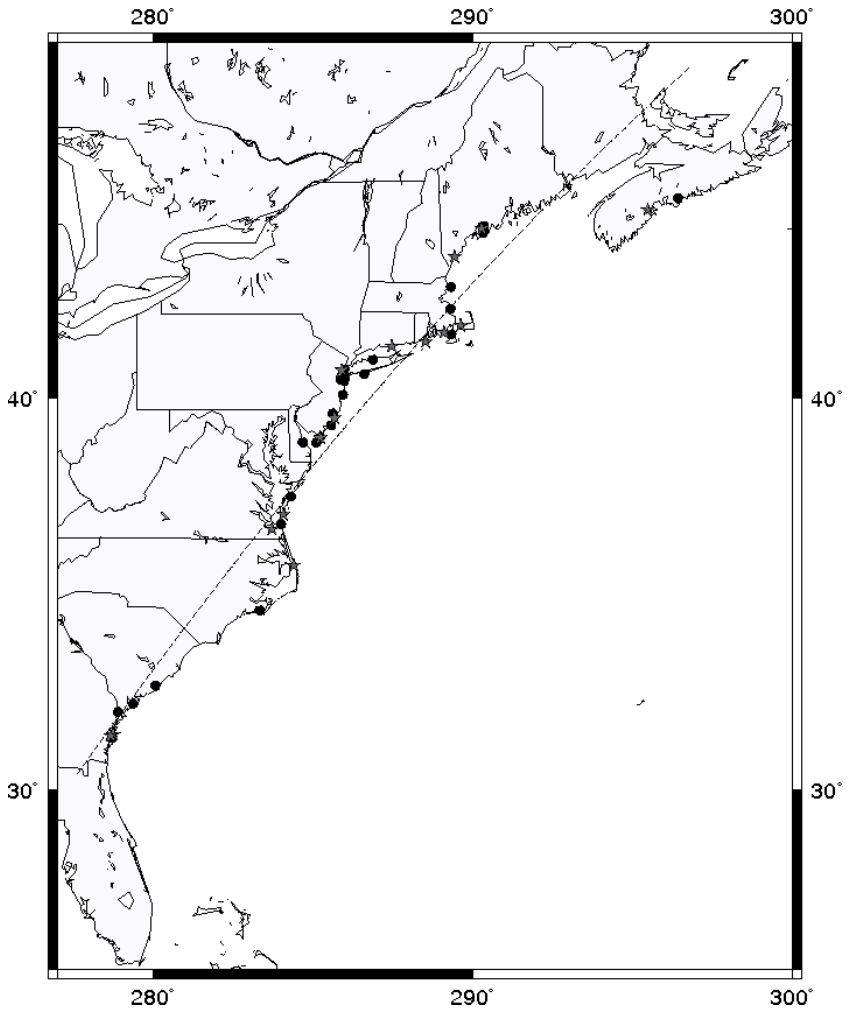


Figure 10.4: Sampled populations of *Fundulus heteroclitus* along the Atlantic coast of North America ($\star$: microsatellite, $\bullet$: allozyme, ---: best-fitting great circle).

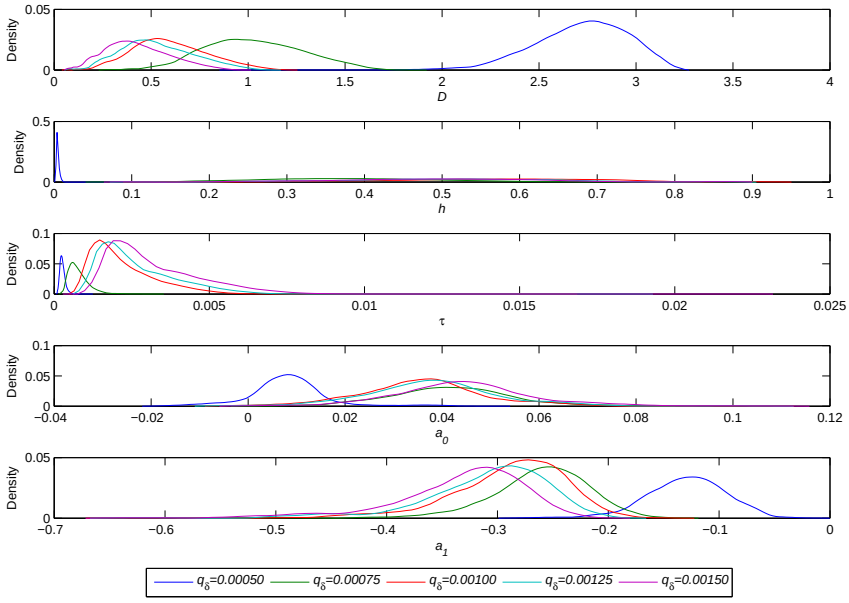


Figure 10.5: Posterior distributions of dispersal and selection parameters for *F. heteroclitus* data when using observed allele frequencies.

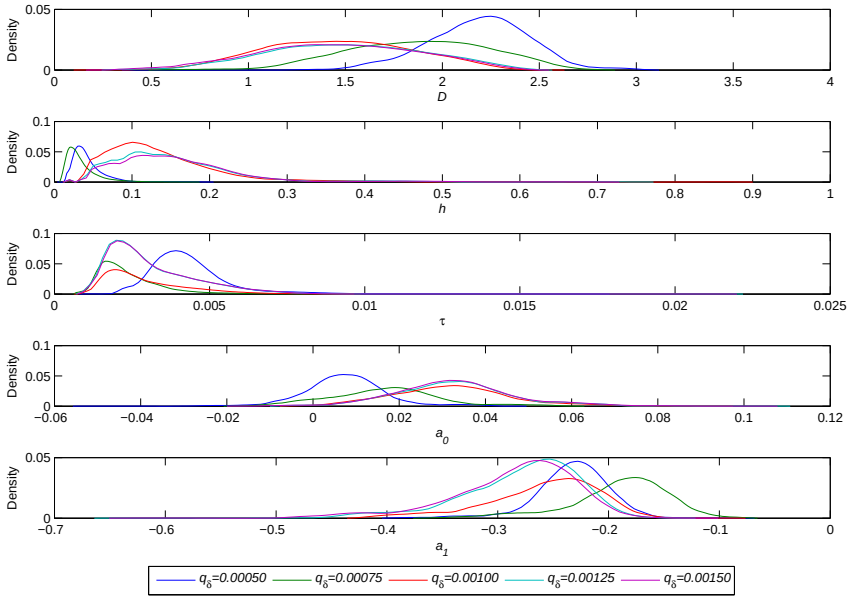


Figure 10.6: Posterior distributions of dispersal and selection parameters for *F. heteroclitus* data when using smoothing splines of observed allele frequencies.

10.8 Appendix I: Numerical solution for the PDE

We compute approximate solutions for equation (10.3) following the numerical method described below. We first scale the PDE in both space and time so that $\Omega = (0, 1)$ and $t \in (0, 1)$ and, thus, respectively substitute $\frac{DT_{\max}}{L^2}$ and $s_{\max}T_{\max}$ for D and $s_{\max}$. Time is discretized into segments of length Δt and the continuous habitat into elements of length Δx (taking into account the curvature of the Earth). The numerical solution of (10.3) is computed by applying a Strang splitting scheme (Strang 1968, Simpson and Landman 2005). At each time step, $p(x, t + \Delta t)$ is computed from $p(x, t)$ by

1. Solving $\frac{\partial p^\star}{\partial t} = \mathcal{G}(p^\star)$ in $\Omega \times (t, t + \frac{\Delta t}{2})$ with initial conditions $p^\star(x, t) = p(x, t)$
2. Solving $\frac{\partial p^{\star\star}}{\partial t} = \mathcal{L}(p^{\star\star})$ in $\Omega \times (t, t + \Delta t)$ with boundary conditions (10.4) and initial conditions $p^{\star\star}(x, t) = p^\star(x, t + \frac{\Delta t}{2})$
3. Solving $\frac{\partial p}{\partial t} = \mathcal{G}(p)$ in $\Omega \times (t + \frac{\Delta t}{2}, t + \Delta t)$ with initial conditions $p(x, t) = p^{\star\star}(x, t)$

This task is done by using 2nd-order methods so that the overall numerical scheme is 2nd-order accurate in both space and time (Strang splitting and inner steps are 2nd-order accurate). The value of the numerical solution at time t^k at N points x_i is denoted as $p_i^k = p(x_i, t^k)$, these value are stored at each time step in the vector $\boldsymbol{p}^k = (p_1^k, \ldots, p_N^k)^T$.

We describe the two numerical schemes used within our Strang splitting method in the two following paragraphs.

Steps 1. and 3. The local growth part of eqn. (10.3) is integrated with a 2nd-order Runge-Kutta method (Teukolsky *et al* 2007)

1. $\boldsymbol{h}_1 = G(\boldsymbol{p}^k)$,
2. $\boldsymbol{h}_2 = G(\boldsymbol{p}^k + \frac{\boldsymbol{h}_1}{2}\delta t)$,
3. $\boldsymbol{p}^{k+1} = \boldsymbol{p}^k + \boldsymbol{h}_2\delta t$

where $G(\boldsymbol{p}) = (\mathcal{G}(p_1), \ldots, \mathcal{G}(p_N))^T$. Note that In our Strang splitting scheme, the time step is $\delta t = \frac{\Delta}{2}$.

Step 2. The Crank-Nicholson discretization of the diffusive part of eqn. (10.3) is

$$\frac{p_i^{k+1} - p_i^k}{\Delta t} = \frac{D}{2}\left(\frac{p_{i-1}^{k+1} - 2p_i^{k+1} + p_{i+1}^{k+1}}{(\Delta x)^2} + \frac{p_{i-1}^k - 2p_i^k + p_{i+1}^k}{(\Delta x)^2}\right) \quad i = 1, \ldots, N \tag{10.8}$$

and boundary conditions (10.4)

$$\frac{p_{i+1}^k - p_{i-1}^k}{2\Delta x} = 0 \quad i = 1, N$$

Thus, introducing $\lambda = \frac{D\Delta t}{(\Delta x)^2}$, we solve

$$\begin{pmatrix} 1+\lambda & -\lambda & 0 & \dots & & 0 \\ -\frac{\lambda}{2} & \ddots & -\frac{\lambda}{2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{\lambda}{2} & \ddots & -\frac{\lambda}{2} \\ 0 & & \dots & 0 & -\lambda & 1+\lambda \end{pmatrix} \boldsymbol{p}^{k+1} = \begin{pmatrix} 1-\lambda & \lambda & 0 & \dots & & 0 \\ \frac{\lambda}{2} & \ddots & \frac{\lambda}{2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\lambda}{2} & \ddots & \frac{\lambda}{2} \\ 0 & & \dots & 0 & \lambda & 1-\lambda \end{pmatrix} \boldsymbol{p}^k \quad (10.9)$$

10.9 Appendix II: Details for the local-linear regression

The weights for the local-linear regression are stored in the diagonal matrix

$$W = Diag(K_\delta(\|\boldsymbol{p}_{(j)} - \boldsymbol{p}\|)) \quad (10.10)$$

where K_δ is the Epanechnikov kernel

$$K_\delta(\|\boldsymbol{p}_{(j)} - \boldsymbol{p}\|) \begin{cases} \propto \delta^{-1}\left(1 - \left(\frac{\|\boldsymbol{p}_{(j)} - \boldsymbol{p}\|}{\delta}\right)^2\right) & \text{if } \|\boldsymbol{p}_{(j)} - \boldsymbol{p}\| \leq \delta \\ = 0 & \text{otherwise} \end{cases} \quad (10.11)$$

and δ is a quantile of the empirical distribution function of the simulated $\|\boldsymbol{p}_{(j)} - \boldsymbol{p}\|$.

Some elements of the parameter vector $\boldsymbol{\phi}$, namely D, h and τ, are constrained (while the others, the a_rs, are unbounded). Thus we must transform these elements so that they become unconstrained before performing regression analysis:

– for $D \in (0, D_{\max})$ and $h \in (0,1)$, i.e. $\phi \in (a,b)$, we set

$$\varphi = \log\left(\frac{\phi - a}{b - \phi}\right) \Leftrightarrow \phi = a + \frac{b-a}{1+\exp(-\varphi)} \tag{10.12}$$

– for $\tau > 0$, i.e. $\phi \in (0, +\infty)$, we set

$$\varphi = \log\phi \Leftrightarrow \phi = \exp\varphi \tag{10.13}$$

Then the vector $\boldsymbol{\varphi}_{(j)}$ contains response variables for each simulated pair.

Explanatory variables are stored in the matrix

$$X = \begin{pmatrix} 1 & (\boldsymbol{p}_{(1)} - \boldsymbol{p})^T \\ \vdots & \vdots \\ 1 & (\boldsymbol{p}_{(n)} - \boldsymbol{p})^T \end{pmatrix} \tag{10.14}$$

and the regression coefficients in

$$\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \dots & \alpha_p \\ \beta_{11} & \dots & \beta_{1p} \\ \vdots & & \vdots \\ \beta_{N1} & \dots & \beta_{Np} \end{pmatrix} \tag{10.15}$$

The least-squares estimate for the regression coefficients is

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = (X^T W X)^{-1} X^T W \boldsymbol{\varphi} \tag{10.16}$$

We form an approximate random sample by setting

$$\boldsymbol{\varphi}_j^\star = \boldsymbol{\varphi}_j - (\boldsymbol{p}^{(j)} - \boldsymbol{p})^T \hat{\boldsymbol{\beta}}$$

and by applying the corresponding inverse transformation for constrained parameters in order to obtain $\boldsymbol{\phi}_j^\star$.

Finally the distribution $(\boldsymbol{\phi}_j^\star, K_\delta(\|\boldsymbol{p}_{(j)} - \boldsymbol{p}\|))$ is used for posterior esitmation purposes (mean, mode, density, HPDI, ...).

10.10 Appendix III: Smoothing splines

Smoothing splines are numerical tools for fitting noisy data, $\tilde{\boldsymbol{y}} = (\tilde{y}_i)$, by a piecewise polynomial function y. Basically spline functions are used for interpolation purposes that can lead to wavy fits when the data are very noisy. The spline

interpolants of sampled data with coordinates $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ cancel the distance between the data and the interpolant

$$\sum (\tilde{y}_i - y(\tilde{x}_i))^2 \tag{10.17}$$

where $\tilde{\boldsymbol{x}} = (\tilde{x}_i)$ are sampled points.

Smoothing splines generalize interpolant splines by introducing a roughness penalty to avoid wavy behaviour of spline interpolation. They are obtained by minimizing

$$\int |y''(x)|^2 \, dx \tag{10.18}$$

which leads to the linear least square estimates of the data.

The roughness penalty for smoothing splines is weighted by a smoothness parameter, λ, that controls the trade-off between data fitting and smoothing. Given $\lambda \geq 0$, the smoothing spline of the data is obtained by minimizing

$$\sum (\tilde{y}_i - y(\tilde{x}_i))^2 + \lambda \int |y''(x)|^2 \, dx \tag{10.19}$$

which leads to a compromise between data fitting and linear least squares.

The smoothness parameter, λ, can be choosen arbitrarily or computed by searching an optimal value for this parameter. In the later case, the popular leave-one-out approach uses the risk in order to choose λ, i.e. finding λ that minimizes

$$R(\lambda) = \sum (\tilde{y}_i - y_{-i,\lambda}(x_i))^2 \tag{10.20}$$

where $y_{-i,\lambda}$ is the smoothing spline obtained when removing observation i from the sampled data for a given value of λ.

Quatrième partie

Discussion générale

L'objectif de cette thèse était d'évaluer l'influence de l'environnement sur la diversité génétique à travers des processus évolutifs. En particulier, il s'agissait de développer des modèles et des méthodes qui permettraient d'estimer l'influence des facteurs environnementaux sur la migration et la sélection naturelle.

La première partie de cette étude portait sur l'estimation des taux de migration récente et de l'influence des facteurs environnementaux sur les flux de gènes. Deux approches ont été étudiées, l'une préexistante, l'autre développée durant la thèse. Les résultats obtenus à partir de données simulées indiquent que les deux méthodes étudiées sont performantes lorsque la différentiation génétique est suffisament élevée ($F_{ST} \geq 0,05$). Par ailleurs une application à des données de génétique humaine a permis de mettre en évidence l'influence de l'altitude sur les migrations entre populations pakistanaises.

La deuxième partie de ce travail portait sur l'influence de gradients environnementaux sur l'adaptation locale. Une méthode d'estimation a été développée à partir de la théorie des clines génétiques sous les effets conjoints de la migration et de la sélection naturelle. L'idée de base était de modéliser les variations spatiales des coefficient de sélection en fonction de gradients environnementaux. L'étude de sensibilité de la méthode développée, à partir de données simulées, a montré que les paramètres de sélection et de dispersion sont correctement estimés lorsque le gradient environemental est clairement établi. L'application au poisson Mummichog a permis de mesurer l'influence de la latitude sur les fréquences alléliques d'un allozyme.

La réponse à la question « Quels facteurs environnementaux influencent tel ou tel processus évolutif ? » est fondamentale dans de nombreuses disciplines. En biologie de la conservation, par exemple, une meilleure compréhension des relations entre environnement et diversité génétique permet de concevoir des stratégies de gestion adaptées. Dans d'autres champs d'application, l'étude de la structuration spatiale de la diversité génétique permet de mettre en évidence l'histoire évolutive d'une espèce et, éventuellement, d'expliquer l'adaptation locale et/ou la propagation de mutations avantageuses.

L'approche utilisée pour répondre à la question biologique posées ci-dessus repose sur les modèles bayésiens hiérarchiques et sur les techniques d'estimation associées. En effet, le formalisme bayésien s'avèrent particulièrement adapté lorsqu'il s'agit de modéliser des processus évolutifs et de relier les facteurs environnementaux aux données génétiques. Bien que ce type d'approches devienne de plus en plus populaire en génétique des populations, la complexité des modèles et des techniques d'estimation requièrent une certaine expertise.

Les méthodes bayésiennes telles que celles qui ont été présentées dans cette thèse sont des outils puissants de la génétique des populations. Les estimations qu'elles produisent apportent d'importantes informations et trouvent des applica-

tions dans différents domaines de la génétique (e.g. biologie de la conservation, évolution, écologie, ...). Cependant les utilisateurs de ce type d'approches doivent être conscients des points développés ci-après pour ne pas utiliser les logiciels qui les implémentent comme des boîtes noires.

Les méthodes bayésiennes utilisent des modèles mathématiques dont les hypothèses ont toutes les chances d'être violées dans le cas des populations naturelles. De manière générale, les principales hypothèses concernent le régime de reproduction (panmixie, consanguinité), l'histoire démographique ou les scénarios évolutifs. Si l'espèce étudiée ne vérifie pas certaines de ces hypothèses, il est probable que les estimations obtenues soient biaisées. Or les articles qui présentent les méthodes n'évaluent que très rarement la robustesse des modèles (dont l'étude ne peut être exhaustive). Il appartient donc à l'utilisateur d'être prudent dans l'interprétation de ses résultats lorsqu'il sait que les hypothèses du modèle ne sont pas vérifiées.

Les études de sensibilité des méthodes et l'évaluation de leur performance donnent une idée de la qualité des estimations produites. Généralement, les articles qui introduisent les méthodes utilisent des données simulées pour étudier les possibilités de ces approches. Par ailleurs, des études complémentaires permettent d'identifier les régions de l'espace des paramètres pour lesquelles la méthode produit des estimations fiables. De telles analyses éclairent les utilisateurs sur les potentialités des méthodes et les aident à interpréter leurs résultats.

Pour les utilisateurs des méthodes, la conclusion est qu'il faut lire attentivement les articles qui présentent ces méthodes, ceux qui les évaluent et le manuel des logiciels qui les implémentent. Par ailleurs, des logiciels de simulation de données (e.g. EASYPOP, Balloux 2001 ; SPLATCHE, Currat et al. 2004) permettent aux utilisateurs de mener leur propre analyse de sensibilité pour tester leur hypothèse. Ils peuvent ainsi générer des données selon des scenarios probables pour les populations qu'ils étudient et examiner les performances de la méthode sous ces conditions.

Les méthodes qui permettent d'obtenir les estimations doivent par ailleurs être connues des utilisateurs. En effet, l'utilisation des méthodes MCMC ou ABC est répandues en génétique des populations, il est donc nécessaire d'en connaître les principes pour les utiliser de façon appropriée. Or l'expérience de cette thèse montre certaines lacunes de ce point de vue dans la mesure où de nombreux utilisateurs analyses leur données sans prendre garde aux réglages des méthodes d'estimation. Si la collecte des données requiert beaucoup de temps et d'énergie, il en est de même pour leur analyse.

Du point de vue du développeur, les méthodes utilisées estiment de plus en plus de paramètres, selon des scénarios de plus en plus complexes. Bien que la puissance de calcul se soit développée, la mise au point et la validation des méthodes d'estimation est un travail qui peut s'avérer long et fastidieux. En parti-

culier l'élaboration des stratégies d'exploration des méthodes MCMC ou le choix de statistiques descriptives dans les approches ABC sont des étapes cruciales qui requierent de nombreux tests.

Pour les raisons citées ci-dessus, il appartient aux personnes qui développent les méthodes de fournir aux utilisateurs toutes les informations nécessaires à une utilisation éclairée de leur approche, pour que les gestionnaires, généticiens et écologues puissent tirer un maximum d'information de leurs données.

Bibliographie

[Abdo et al., 2004] Abdo, Z., Crandall, K. A., and Joyce, P. (2004). Evaluating the performance of likelihood methods for detecting population structure and migration. *Mol Ecol*, 13(4) :837–851.

[Adams et al., 2006] Adams, S. M., Lindmeier, J. B., and Duvernell, D. D. (2006). Microsatellite analysis of the phylogeography, pleistocene history and secondary contact hypotheses for the killifish, fundulus heteroclitus. *Mol Ecol*, 15(4) :1109–1123.

[Avise, 2004] Avise, J. (2004). *Molecular Markers, Natural History, and Evolution, 2nd edn.* Sinauer Associates, Sunderland, Massachusetts, 2nd edition.

[Bahlo and Griffiths, 2000] Bahlo, M. and Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor Popul Biol*, 57(2) :79–95.

[Balding and Nichols, 1995] Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2) :3–12.

[Balding and Nichols, 1997] Balding, D. J. and Nichols, R. A. (1997). Significant genetic correlations among caucasians at forensic dna loci. *Heredity*, 78(6) :583–589.

[Ballain et al., 1998] Ballain, T., Litaudon, P., Martiel, J. L., and Cattarelli, M. (1998). Role of the net architecture in piriform cortex activity : analysis by a mathematical model. *Biol Cybern*, 79(4) :323–336.

[Balloux, 2001] Balloux, F. (2001). Easypop (version 1.7) : a computer program for population genetics simulations. *Journal of Heredity*, 92(3) :301–302.

[Barton and Hewitt, 1985] Barton, N. and Hewitt, G. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, 16 :113–6148.

[Beaumont and Balding, 2004] Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol*, 13(4) :969–980.

[Beaumont and Rannala, 2004] Beaumont, M. A. and Rannala, B. (2004). The bayesian revolution in genetics. *Nat Rev Genet*, 5(4) :251–261.

[Beerli and Felsenstein, 2001] Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A*, 98(8) :4563–4568.

[Bossart and Pashley Prowell, 1998] Bossart, J. and Pashley Prowell, D. (1998). Genetic estimates of population structure and gene flow : Limitations, lessons and new directions. *Trends in Ecology & Evolution*, 13(5) :202–206.

[Brooks, 1998] Brooks, S. P. (1998). Markov chain monte carlo method and its application. *Journal of the Royal Statistical Society Series D - The Statistician*, 47(1) :69–100.

[Brooks et al., 2003] Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 65(1) :3–39.

[Brown and Chapman, 1991] Brown, B. L. and Chapman, R. W. (1991). Gene flow and mitochondrial dna variation in the killifish, fundulus heteroclitus. *Evolution*, 45 :1147–1161.

[Cann et al., 2002] Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L. (2002). A human genome diversity cell line panel. *Science*, 296(5566) :261–262.

[Carmichael et al., 2001] Carmichael, L. E., Nagy, J. A., Larter, N. C., and Strobeck, C. (2001). Prey specialization may influence patterns of gene flow in wolves of the canadian northwest. *Mol Ecol*, 10(12) :2787–2798.

[Cashon et al., 1981] Cashon, R. E., Beneden, R. J. V., and Powers, D. A. (1981). Biochemical genetics of fundulus heteroclitus (l.). iv. spatial variation in gene frequencies of idh-a, idh-b, 6-pgdh-a, and est-s. *Biochem Genet*, 19(7-8) :715–728.

[Charbonnel et al., 2002a] Charbonnel, N., Angers, B., Rasatavonjizay, R., Bremond, P., Debain, C., and Jarne, P. (2002a). The influence of mating system, demography, parasites and colonization on the population structure of biomphalaria pfeifferi in madagascar. *Mol Ecol*, 11(11) :2213–2228.

[Charbonnel et al., 2002b] Charbonnel, N., Quesnoit, M., Razatavonjizay, R., Brémond, P., and Jarne, P. (2002b). A spatial and temporal approach to microevo-

lutionary forces affecting population biology in the freshwater snail biomphalaria pfeifferi. *Am Nat*, 160(6) :741–755.

[Corander and Marttinen, 2006] Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Mol Ecol*, 15(10) :2833–2843.

[Corander et al., 2004] Corander, J., Waldmann, P., Marttinen, P., and Sillanpää, M. J. (2004). Baps 2 : enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, 20(15) :2363–2369.

[Craven and Wahba, 1979] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31 :377–403.

[Currat et al., 2004] Currat, M., Ray, N., and Excoffier, L. (2004). splatche : a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, 4(1) :139–142.

[Evanno et al., 2005] Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study. *Molecular Ecology*, 14(8) :2611–2620.

[Excoffier and Smouse, 1994] Excoffier, L. and Smouse, P. E. (1994). Using allele frequencies and geographic subdivision to reconstruct gene trees within a species : molecular variance parsimony. *Genetics*, 136(1) :343–359.

[Falush et al., 2003] Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics*, 164(4) :1567–1587.

[Faubet and Gaggiotti, 2008] Faubet, P. and Gaggiotti, O. E. (2008). A new bayesian method to identify the environmental factors that influence recent migration. *Genetics*, 178(3) :1491–1504.

[Faubet et al., 2007] Faubet, P., Waples, R. S., and Gaggiotti, O. E. (2007). Evaluating the performance of a multilocus bayesian method for the estimation of migration rates. *Mol Ecol*, 16(6) :1149–1166.

[Felsenstein, 1975] Felsenstein, J. (1975). Genetic drift in clines which are maintained by migration and natural selection. *Genetics*, 81(1) :191–207.

[Ficetola et al., 2007] Ficetola, G. F., Garner, T. W. J., and Bernardi, F. D. (2007). Genetic diversity, but not hatching success, is jointly affected by postglacial colonization and isolation in the threatened frog, rana latastei. *Mol Ecol*, 16(9) :1787–1797.

[Fife and Peletier, 1981] Fife, P. C. and Peletier, L. A. (1981). Clines induced by variable selection and migration. *Proceedings of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, 214(1194) :99–123.

[Fisher, 1937] Fisher, R. A. (1937). The wave of advance of advantegeous genes. *Ann Eugenics*, 7 :355–369.

[Foll and Gaggiotti, 2006] Foll, M. and Gaggiotti, O. (2006). Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, 174(2) :875–891.

[François et al., 2006] François, O., Ancelet, S., and Guillot, G. (2006). Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2) :805–816.

[Freville et al., 2001] Freville, H., Justy, F., and Olivieri, I. (2001). Comparative allozyme and microsatellite population structure in a narrow endemic plant species, centaurea corymbosa pourret (asteraceae). *Mol Ecol*, 10(4) :879–889.

[Gaggiotti, 2006] Gaggiotti, O. E. (2006). Evolutionary population genetics : Were the vikings immune to hiv ? *Heredity*, 96(4) :280–281.

[Gaggiotti et al., 2004] Gaggiotti, O. E., Brooks, S. P., Amos, W., and Harwood, J. (2004). Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Molecular Ecology*, 13(4) :811–825.

[Gaggiotti et al., 2002] Gaggiotti, O. E., Jones, F., Lee, W. M., Amos, W., Harwood, J., and Nichols, R. A. (2002). Patterns of colonization in a metapopulation of grey seals. *Nature*, 416(6879) :424–427.

[Gelman et al., 1995] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[Geyer, 1991] Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood. In Keramidas EM , editor, *Computing Science and Statistics : Proceedings of the 23rd Symposium on the Interface*, pages 156–163, Interface Foundation, Fairfax Station,Virginia.

[Giordano et al., 2007] Giordano, A. R., Ridenhour, B. J., and Storfer, A. (2007). The influence of altitude and topography on genetic structure in the long-toed salamander (ambystoma macrodactulym). *Mol Ecol*, 16(8) :1625–1637.

[Golightly and Wilkinson, 2006] Golightly, A. and Wilkinson, D. J. (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *J Comput Biol*, 13(3) :838–851.

[Green and Silverman, 1994] Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach*. Chapman and Hall, London.

[Green, 1995] Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4) :711–732.

[Guillot et al., 2005] Guillot, G., Estoup, A., Mortier, F., and Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics*, 170(3) :1261–1280.

[Hastings, 1993] Hastings, A. (1993). Complex interactions between dispersal and dynamics - lessons from coupled logistic equations. *Ecology*, 74(5) :1362–1372.

[Hudson et al., 1992] Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992). Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2) :583–589.

[Huxley, 1939] Huxley, J. (1939). Clines : an auxiliary method in taxonomy. *Bijdragen tot de Dierkunde (Leiden)*, 27 :491–520.

[Keller, 1984] Keller, J. (1984). Genetic variability due to geographical inhomogeneity. *Journal of Mathematical Biology*, 20(2) :223–230.

[Kingman, 1982a] Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes and their Applications*, 13(3) :235–248.

[Kingman, 1982b] Kingman, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Probability*, 19A :19–27.

[Lynch and Crease, 1990] Lynch, M. and Crease, T. J. (1990). The analysis of population survey data on dna sequence variation. *Mol Biol Evol*, 7(4) :377–394.

[May et al., 1975] May, R. M., Endler, J. A., and McMurtrie, R. E. (1975). Gene frequency clines in the presence of selection opposed by gene flow. *The American Naturalist*, 109(970) :659.

[McRae, 2006] McRae, B. H. (2006). Isolation by resistance. *Evolution*, 60(8) :1551–1561.

[Moody, 1979] Moody, M. (1979). Polymorphism with migration and selection. *Journal Of Mathematical Biology*, 8 :73–109.

[Moody, 1981] Moody, M. (1981). Polymorphism with selection and genotype-dependent migration. *Journal Of Mathematical Biology*, 11 :245–267.

[Nagylaki, 1976] Nagylaki, T. (1976). Clines with variable migration. *Genetics*, 83(4) :867–886.

[Nagylaki, 1977] Nagylaki, T. (1977). *Lecture Notes in Biomathematics 15*. Springer-Verlag, Berlin.

[Nagylaki, 1978a] Nagylaki, T. (1978a). Clines with asymmetric migration. *Genetics*, 88(4) :813–827.

[Nagylaki, 1978b] Nagylaki, T. (1978b). Random genetic drift in a cline. *Proc Natl Acad Sci U S A*, 75(1) :423–426.

[Nagylaki, 1989] Nagylaki, T. (1989). *Models in Population Biology*, chapter The diffusion model for migration and selection, pages 55–75. The American Mathematical Society, Providence, Rhode Island.

[Nagylaki and Moody, 1980] Nagylaki, T. and Moody, M. (1980). Diffusion model for genotype-dependent migration. *Proc Natl Acad Sci U S A*, 77(8) :4842–4846.

[Nei, 1973] Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A*, 70(12) :3321–3323.

[Nielsen and Wakeley, 2001] Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation : a markov chain monte carlo approach. *Genetics*, 158(2) :885–896.

[Novembre et al., 2005] Novembre, J., Galvani, A. P., and Slatkin, M. (2005). The geographic spread of the ccr5 delta32 hiv-resistance allele. *PLoS Biol*, 3(11) :e339.

[Ohta, 1982] Ohta, T. (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc Natl Acad Sci U S A*, 79(6) :1940–1944.

[Paetkau et al., 1995] Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C. (1995). Microsatellite analysis of population structure in canadian polar bears. *Mol Ecol*, 4(3) :347–354.

[Peletier, 1978] Peletier, L. (1978). A nonlinear eigenvalue problem occuring in population genetics. In Beniland, P. and Robert, J., editors, *Journées d'Analyse Non Linéaire*, pages 170–187.

[Powers and Place, 1978] Powers, D. A. and Place, A. R. (1978). Biochemical genetics of fundulus heterolitus (l.). i. temporal and spatial variation in gene frequencies of ldh-b, mdh-a, gpi-b, and pgm-a. *Biochem Genet*, 16(5-6) :593–607.

[Powers et al., 1986] Powers, D. A., Ropson, I., Brown, D. C., Van Beneden, R., Cashon, R., Gonzalez-Villasenor, L. I., and Dimichele, J. A. (1986). Genetic variation in fundulus heteroclitus : Geographic distribution. *Amer. Zool.*, 26(1) :131–144.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–959.

[Qamar et al., 2002] Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C., and Mehdi, S. Q. (2002). Y-chromosomal dna variation in pakistan. *Am J Hum Genet*, 70(5) :1107–1124.

[Quintana-Murci et al., 2004] Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., Rengo, C., Al-Zahery, N., Semino, O., Santachiara-Benerecetti, A. S., Coppa, A., Ayub, Q., Mohyuddin, A., Tyler-Smith, C., Mehdi, S. Q., Torroni, A., and McElreavey, K. (2004). Where west meets east : the complex mtdna landscape of the southwest and central asian corridor. *Am J Hum Genet*, 74(5) :827–845.

[Ramsay et al., 2007] Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations : a generalized smoothing approach. *Journal Of The Royal Statistical Society Series B*, 69(5) :741–796.

[Rannala and Mountain, 1997] Rannala, B. and Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proc Natl Acad Sci U S A*, 94(17) :9197–9201.

[Robert, 1994] Robert, C. P. (1994). *The Bayesian Choice : A Decision-Theoretic Motivation*. Springer, New York.

[Ropson et al., 1990] Ropson, I. J., Brown, D. C., and Powers, D. A. (1990). Biochemical genetics of fundulus heteroclitus (l.) vi. geographical variation in the gene frequencies of 15 loci. *Evolution*, 44 :16–26.

[Rosenberg et al., 2002] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602) :2381–2385.

[Rousset, 1996] Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142(4) :1357–1362.

[Slatkin, 1973] Slatkin, M. (1973). Gene flow and selection in a cline. *Genetics*, 75(4) :733–756.

[Slatkin, 1987] Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236(4803) :787–792.

[Slatkin, 1995] Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1) :457–462.

[Slatkin and Maruyama, 1975] Slatkin, M. and Maruyama, T. (1975). Genetic drift in a cline. *Genetics*, 81(1) :209–222.

[Spiegelhalter et al., 2002] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measure of model complexity and fit. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(4) :583–616.

[Vitalis et al., 2001] Vitalis, R., Dawson, K., and Boursot, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics*, 158(4) :1811–1823.

[Wahba, 1985] Wahba, G. (1985). A comparison of gcv and gml for choosing the smoothness parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 13 :1378–1402.

[Waples and Gaggiotti, 2006] Waples, R. S. and Gaggiotti, O. (2006). What is a population ? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15(6) :1419–1439.

[Weir and Cockerham, 1984] Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population-structure. *Evolution*, 38(6) :1358–1370.

[Wessel and Smith, 1998] Wessel, P. and Smith, W. H. F. (1998). New, improved version of generic mapping tools released. *Eos Trans. AGU*, 79 :579.

[Whitlock and McCauley, 1999] Whitlock, M. C. and McCauley, D. E. (1999). Indirect measures of gene flow and migration : Fst not equal to 1/(4nm + 1). *Heredity*, 82 (Pt 2) :117–125.

[Wikle and Hooten, 2006] Wikle, C. K. and Hooten, M. B. (2006). *Hierarchical Modelling for the Environmental Sciences*, chapter Hierarchical Bayesian spatio-temporal models for population spread. Oxford University Press.

[Wilson and Rannala, 2003] Wilson, G. A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, 163(3) :1177–1191.

[Wright, 1951] Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15 :323–354.

Cinquième partie

Annexe

Estimation de la fréquence d'un allèle

Les mises à jour de p et de α utilise respectivement l'échantillonneur de Gibbs et l'algorithme MH.

Mise à jour de p L'expression de la loi conditionnelle de p montre que la fréquence de l'allèle suit une loi de type Beta, i.e.

$$p|\alpha, k \backsim \beta(\alpha + k, \alpha + 2N - k) \tag{21}$$

Mise à jour de α Les nouvelles valeurs pour le paramètre α sont proposoées selon une loi log-normale centrée en $\log \alpha$, i.e.

$$\log \alpha' \backsim \mathcal{N}(\log \alpha, \sigma_\alpha^2) \tag{22}$$

avec $\sigma_\alpha^2 = 1$

La transition de α vers α' est accepté avec la probabilité

$$\alpha(\alpha, \alpha') = \min\left(1, \frac{\Gamma(2\alpha')\Gamma(\alpha)^2}{\Gamma(2\alpha)\Gamma(\alpha')^2}(p(1-p))^{\alpha'-\alpha} \exp\left(-\frac{(\log \alpha')^2 - (\log \alpha)^2)}{2}\right)\right) \tag{23}$$